中国人工智能系列白皮书

——多语种智能信息处理

中国人工智能学会

二〇二二年六月

《中国人工智能系列白皮书》编委会

主　　任：戴琼海

执行主任：王国胤

副 主 任：陈　杰　刘成林　刘　宏　孙富春　王恩东　王文博
赵春江　周志华

委　　员：班晓娟　曹　鹏　陈　纯　陈松灿　邓伟文　董振江
杜军平　付宜利　古天龙　桂卫华　何　清　胡国平
黄河燕　季向阳　贾英民　焦李成　李　斌　刘　民
刘庆峰　刘增良　鲁华祥　马华东　苗夺谦　潘　纲
朴松昊　钱　锋　乔俊飞　孙长银　孙茂松　陶建华
王卫宁　王熙照　王　轩　王蕴红　吾守尔·斯拉木
吴晓蓓　杨放春　于　剑　岳　东　张小川　张学工
张　毅　章　毅　周国栋　周鸿祎　周建设　周　杰
祝烈煌　庄越挺

《中国人工智能系列白皮书----多语种智能信息处理》编写组

组　　长：吾守尔·斯拉木

执行组长：张华平　骆　曦

组　　员：阿拉坦巴根那　艾斯卡尔·艾木都拉　布音其其格
陈　诗　陈自岩　飞　龙　高　歌　高光来　高　璐
哈妮克孜伊拉洪　哈　斯　蒋盛益　金　罡　赖　文

李　琳　林　民　林楠铠　刘江维　刘　群　刘　瑞
刘维锋　玛依热·依布拉音　米吉提·阿不里米提
娜仁高娃　诺明花　丘心颖　申影利　史云伟
苏向东　孙　晓　吐尔地托合提　万福成　王连喜
王斯日古楞　王炜华　王一帆　汪张龙　王治敏
武　智　肖莉娴　徐豪杰　徐　悦　严若豪　杨冰冰
杨曼芝　杨子妍　姚　洲　于洪志　袁亮杰　张宝华
张　晖　张卫国　张霄军　张新猛　赵慧周　赵小兵
赵旭阳　周毛克　宗　浩

目　录

第 1 章 多语种智能信息处理概述

1.1 多语种信息处理的必要性

“语言是了解一个国家最好的钥匙”，我国的“一带一路”重大战略决策和民族政策都将文化交流作为重要内容，习总书记在中国 - 东盟 30 周年纪念峰会上强调：要共建友好家园，深化文明交流互鉴，使双方民众更加相知、相亲、相融。自然语言处理（NLP）技术是促进文化交流的重要技术手段，通过 NLP 技术可以提高交流和沟通能力，对我国战略和政策的推进很有帮助。

现有的 NLP 技术在主流语言已经取得了很好效果，特别是汉语和英语方面，人工智能算法在某些领域已经超越了人类的表现。但是在小语种领域，相关分析算法效果不佳，对真实场景的分析效果差，不能很好的在实际应用中发挥作用，对我国的政策推进的支持有限。在此形势下，对小语种领域 NLP 技术的研究就尤为重要。

当今世界正处于百年未有之大变局，人工智能、量子技术、虚拟现实等新一代人工智能技术正在深刻改变人类的生产和生活方式，把握数字网、网络化、智能化融合发展的契机，开展多语言混合智能信息处理领域的相关研究，对于发挥我国多民族、多语言的优势与特色，促进我国“一带一路”沿线国家语言和信息的互联互通意义重大。

1.2 多语种信息处理现状

在多语种信息处理方面，本文将分别从底层操作系统构建，多语种信息处理平台以及后期应用方面进行介绍。

1.2.1 国产多语种操作系统和信创软件

操作系统软件作为信息技术的基础性系统软件平台，能否实现国产替代，直接影响着我国互联网生态的自主可控，习近平总书记多次强调“没有网络安全就没有国家安全”、“不掌握核心技术，我们就会被卡脖子、牵鼻子，不得不看别人脸色行事”、“关键核心技术要不来、

买不来、讨不来”，操作系统这个最基础、最底层的软件很可能成为我国发展的绊脚石。自 20 世纪 80 年代以来，我国主要少数民族语言信息化的科技工作者们在这方面做了大量工作，30 多年来，基本紧跟了汉文信息化水平，其发展历程主要分为 DOS 操作系统、Windows 操作系统和 Linux 操作系统三个阶段。在此期间，新疆大学、西北民族大学、青海大学、西藏大学、内蒙古大学、广西计算中心等单位分别在各自的领域展开了大量研究工作，多语种的编码标准、键盘布局标准、字体标准相继制定，并根据这些标准先后基于 DOS、Windows、Linux 等操作系统开发了多语种的字库、输入法、字处理软件，以及多语种操作系统软件产品，并在各地区和领域进行了广泛推广，也为后续少数民族语言信息化打下了坚实基础。

近年来，在国家的大力支持下，我国信息技术应用创新产业（简称“信创产业”）取得了巨大成就，国产 CPU 取得群体性突破，达到或接近国际先进水平，实现量产并大批量应用于国产服务器、计算机终端、移动终端等电子产品中，初步打破了美国等西方国家在这一领域的长期垄断；国产基础软件快速集约发展，数据库系统、中间件、办公软件均取得长足进步，大幅缩小了与国外基础软件的差距，开始规模化普及应用，并与基于国产 CPU 的计算机相配套，逐步形成我国自主可控的信创产业生态链。目前，国产桌面操作系统产业方兴未艾，已经兴起的 10 多家操作系统企业主要力量集中在汉文领域，而在我国多民族大家庭中占据重要位置的各少数民族语言领域，由于相对于产出而言投入较大，国内各操作系统企业在此方面的工作投入相对较少，这将制约各少数民族语言操作系统，以及各少数民族信息化的发展，如不尽快开展相关工作，随着时间的推移，少数民族信息化事业将会越发滞后。

少数民族语言文字历史悠久、源远流长，是中华民族文化与文明的重要组成部分。在信息技术应用创新的背景下，基于国产软硬件体

系，构建多语种信息处理技术体系，既可以有效地解决我国少数民族语言信息处理中的大量基础性、共同性的关键和核心问题，避免重复开发，保证各民族语言处理软件的兼容性和相互支持，进而促进少数民族信息化事业的发展，也可以使少数民族语言信息化摆脱受制于国外基础软硬件平台的现状，保障我国语言信息化的安全，促进我国多语种信息处理技术和成果对一带一路建设中的辐射、引领性作用，为国产多语种信创软件面向一带一路的推广奠定基础。为了推进此项工作，2016 年 12 月吾守尔院士团队发起成立了“国产多语种操作系统技术联盟”，2017 年承担了国家语委重大科研项目“国产多语种桌面操作系统通用规范研制”，2021 年加入了信创工委会，并联合相关知名单位大力推进国产多语种操作系统和信创软件研发工作。

1.2.2 多语言互译平台

自习近平总书记提出“一带一路”倡议以来，共建“一带一路”已经成为增进各国民众福祉重大举措，成果正在惠及世界。“一带一路”建设的主要内容是政策沟通、设施联通、贸易畅通、资金融通、民心相通，这“五通”哪一通也离不开“语言相通、信息互通”。“一带一路”沿线国家中有 53 种官方语言，其中大多数为非通用语种，不同语言与汉语之间的语言沟通、信息互通障碍已成为制约“一带一路”各项合作交流的关键问题，也是我国及周边国家反恐维稳和情报舆情分析的主要掣肘。同时也应该注意到，语言的区域性和地缘性特征显著，同语系语言在语言特性及使用上存在一定的相似性，通过整合国内展开机器翻译方面研究的机构和资源，研究同语系及跨语系自然语言的互译工作，将有效促进“一带一路”沿线国家的文化、科技、医疗、教育、旅游等方面的交流与合作。

在多语言互译平台方面，新疆大学构建的丝绸之路经济带多语言互译平台是当前多语言互译平台的代表之一，于 2019 年成为工信部“新一代人工智能产业创新重点任务”智能翻译领域的新疆唯一揭榜

潜力单位。目前，通过采集维吾尔语 - 汉语、汉语 - 维吾尔语、哈萨克语（哈国）- 中文、中文 - 哈萨克语（哈国）、吉尔吉斯语（吉国）- 中文、中文 - 吉尔吉斯语（吉国）、乌尔都语 - 中文、中文 - 乌尔都语、乌兹别克语 - 中文、德语 - 中文、法语 - 中文的语料，建设了以上 11 种翻译的语料数据，使用改进的端到端多语言神经网络模型 Transformer，实现了以上 11 种机器翻译产品。其中，与科大讯飞合作研发的维汉双向语音翻译系统在新疆脱贫攻坚、乡村振兴及 24 万驻村干部的“访惠聚”工作中应用，显著提高了社会治理的能力和水平。为了进一步提升，目前新疆大学仍然在进行以下工作。

1. 构建大规模、多层次综合型多语言知识库、语料库系统

大规模采集并构建中亚、南亚、西亚主要国家及民族的自然语言语料库，重点开展哈萨克语、柯尔克孜语、乌兹别克语、土耳其语与汉语平行语料库、语音数据库建设，研究其语言形态、文字特性等方面的特征，研究多语言知识的挖掘及采集方法和模型，构建大规模、多层次综合型多语言知识库、语料库系统。

2. 多语言智能理解技术研究

向阿拉伯语系、阿尔泰语系语族和印度 - 伊朗语族开展了词法分析、句法分析，以及多语言的词法、句法、语义、篇章、情感、蕴含、信息抽取等语言分析方法等方面的研究与开发工作。研究复杂形态语言和长距离语言模型、跨语言文法推导方法等，根据语言特点采用规则、统计、神经网络等不同方法实现。

3. 跨语系自然语言机器翻译方法和模型研究

研究不同语系语言机器翻译方法、形态复杂语言机器翻译、资源匮乏语言机器翻译、枢轴语言机器翻译等理论与方法，重点突破汉语 - 印度伊朗语族、汉语 - 阿拉伯语智能机器翻译核心技术。

4. 同语系自然语言机器翻译方法和模型研究

利用语言间的相似特性，对汉语 - 中亚西亚阿尔泰语系多种语

言互译技术，主要利用单语和双语数据的神经机器翻译、基于迁移学习的多语言神经机器翻译框架等，开展同语系自然语言机器翻译方面的方法和模型研究。针对低资源语音翻译研究，使用基于迁移学习方法，采用半监督机制提升语音识别的鲁棒性，构建低资源高鲁棒性语音识别系统；利用模型对语音信号文本、说话人、信道分别提取编码变量；基于不同语种的发音机制相同这一假设和数据驱动与知识引导相结合的人工智能新方法，开展基于 Global Phone 的多语种统一声学建模方案研究，提高不同语言的数据共享能力和自动化水平。利用端到端语音识别方法，直接从语音波形映射到识别输出，提高识别效果。

1.2.3 面向公共安全的多语种舆情监测、预警系统

在多语种信息处理的应用方面，以吾守尔院士团队构建的多语种舆情监测、预警系统为例。2020 年 2 月 3 日，习近平总书记在中央政治局常委会会议讲话中，明确指出“要加强舆情跟踪研判，主动发声、正面引导，强化融合传播和交流互动，让正能量始终充盈网络空间”。新疆既是我国反恐维稳的主要阵地，也是境内外敌对势力在意识形态领域进行渗透的主要场地，敌对势力主要使用阿尔泰语系、阿拉伯语系中的少数民族语言在网络上进行渗透和传播暴恐思想，组织暴恐活动，极少使用英语和汉语等语言。目前虽有一些较为成熟的中英文网络舆情管控系统，但缺乏有效的多语言网络舆情管控系统和平台，针对这一现状及新疆多民族、多语言和多元文化的特点，团队首次提出并实现维、哈、柯多语言自动识别、转换及正规化方法，提出并实现维、哈、柯语义分词方法，开展网络文本、语音、视觉等大数据的采集、聚类、情感、异常、预警、可视化等方面的模型、算法及方法研究。开展情报收集、面向反恐的视频与音频分析及检索、人脸识别、说话人识别、目标对象实时跟踪等感知，这对于我国反恐维稳和自治区社会稳定长治久安具有重要意义。

1.3 多语种信息处理展望

2018 年 4 月，中国工程院信息与电子工程学部在北京举办了“丝绸之路经济带多种语言互译平台开发应用研讨会”。出席了 7 位院士，国家部委有关司局领导，以及高校、研究院所和企业人工智能专家共 80 余人。大会通过了《关于加快推进丝绸之路经济带多种语言互译平台开发应用的倡议》，对开发建设丝路多种语言互译平台的重要性、必要性与紧迫性给予充分肯定，对多语种信息处理提出了更高的要求，要求加快推进。2021 年 6 月发起成立了中国人工智能学会多语种智能信息处理专业委员会，旨在凝聚国内知名企业、高校和研究机构构建丝绸之路经济带多语言互译平台，全面提升丝路沿线国家和地区的交通、文化、教育、农业、林业、水利、智慧城市、环保、防灾减灾、公共安全、旅游业等领域的应用研发合作。

国之交在于民亲，民相亲在于心通，一带一路沿线国家涉及的语言多种多样，其中大多数为非通用语言，借助智能化、信息化手段研究同语系和跨语系的机器翻译，使得不同国家、不同民族的人们顺畅沟通和理解，才能够更好地多方面的交流与合作，从而实现“民心相通”。面对百年未有之大变局，把握数字化、网络化、智能化发展机遇，加快推进多语种智能信息处理基础算法、应用平台研发工作，切实提高面向公共安全的舆情监测、预警能力，对于缩小数字鸿沟，推动数字经济转型，构建网络空间命运共同体意义深远。

多语种智能信息处理包括对国内民族语言和其他国家语言的处理，涵盖数据收集、语言分析、多语翻译及多语挖掘等分析步骤，包括词法分析、句法分析、语言模型构建、语音识别等多个关键技术。本书纵观多语种的发展，从多语种的技术、应用和研究机构方面介绍了当前多语种信息智能处理的现状。

本章编写人员：

吾守尔·斯拉木、张华平、张宝华

第 2 章 民族语言智能信息处理

2.1 维吾尔文智能信息处理

近年来，在国内汉语智能信息处理技术研究成果的先导作用下，维吾尔语文本信息的智能处理也得到了长足发展。本章主要以新疆大学研究团队近几年研究工作进展为主，介绍维吾尔文语义串抽取，聚类，自动摘要，维吾尔语本体构建等方面的研究现状，以及正面临的主要问题和未来要开展的主要工作。

2.1.1 维吾尔文语义串抽取

从文字表面上看，维吾尔文是以空格隔开的词序列，在这一特点上与英文类似。因此，常以空格作为自然分隔符进行简单分词。其实，维吾尔文中能表达一个最基本的、具体而完整语义的语言单元，在很多情况下不仅仅是一个以空格隔开的单词，而是它与上下文若干个词的稳定组合。因此，维吾尔文中能表达一个完整语义，或者说在实际语言环境中能充当一个实词的串，可分为以下两类：

（1）单词语义串：是一个维吾尔文单词，是一个无空格字母串，语义完整且独立运用语言单元，可用空格作为自然分隔符切分得到。

（2）多词语义串：是若干个维吾尔文单词的稳定组合，其特点是：语义完整，在真实语言环境中充当一个实词，不能以空格分开。结构稳定，在大规模语料中具有较高的流通度，也是独立运用语言单元。如，“تىك ئۇچار ئايروپىلان”（直升机）。

针对以上情况，一种基于统计和浅层语言分析的维吾尔文语义串快速抽取方法在不同规模的语料展现出很好的效果，也被证明能够应用到维吾尔文文本挖掘多个领域中。主要思路是，采用一种多层动态索引结构为大规模文本建词索引，然后是结合维吾尔文词间关联规则采用一种改进的 n 元递增算法进行词串扩展并发现文本中的可信频繁模式，最终依次判断频繁模式串结构完整性从而得到文本中的语

义串集。

2.1.2 维吾尔文文本聚类

因为单词的语义表达能力有限，从而以词特征的文本表示模型很难发挥学习算法最佳性能，因此用超出词语边界的语言单位——语义串来表示文本仍然是一个研究热点。以语义串为文本特征的维吾尔文文本聚类方法中，以带权语义串集来表示每一个文本，提出一种基于集合的文本相似度度量方法，通过多个实验来验证其正确性和有效性。

1. 文本表示

假设，文本 d_i 中有 n 个语义串（文本主体项）$\{S_1,S_2,\ldots,S_j,\ldots,S_{n-1},S_n\}$，若用 W_j 来表示语义串 S_j 的权重，则可以用一个二元组（S_j ，W_j）来表示每一个，那么文本 d_i 就可以表示成 n 个带权语义串的集合，即 $\{(S_1, W_1), (S_2, W_2), \ldots, (S_j, W_j) ,\ldots, (S_{n-1}, W_{n-1}) , (S_n, W_n)\}$。

计算语义串权重时，主要考虑语义串对于表示文本主题的贡献度。首先，邻接特征量表示语义串在语用环境中的结构完整性，而结构完整的词串总是能表达与文本主题相关的关键信息。除此之外，语义串的长度与其表达的信息量是成正比的关系，因此长度越长，语义串表达的信息量也越大，其语义更具体而完整。因此，给出了如下权重计算公式，即

$$W_j = AE_{weight} \times \sqrt{Unit\_count}$$

其中，W_j 是文本 d_i 中语义串 S_j 的权重，AE_{weight} 是其邻接熵，$Unit\_count$ 是其长度（语义串单词个数）。

2. 文本相似性度量

根据基于语义串的文本表示方法，采取了一种类似于 Jaccard 相似度的文本相似性度量方法。相关术语定义如下：

（1）$D=\{d_i\}$: D 为文本集，d_i 是 D 中第 i 个文本

（2）文本主体项集 T_{di}：是文本集 D 中文本 d_i 的特征集（带权语义串集）。

（3）文本权重：是文本 d_i 的特征集 T_{di} 中全部 n 个特征（语义串）权重之和，即

$$Weight(T_{di}) = \sum_{k=1}^{n} W_k$$

（4）主题项交集 $T_{di} \cap T_{dj}$：是文本 T_{di} 和 T_{dj} 共有的语义串集。

（5）主题项交集权重 $Weight(T_{di} \cap T_{dj})$：是两个文本相交项集全部语义串权重之和。

两个文本间的相似性是这两个文本在主题上的共性，用主题项（语义串）集来表示文本，那么两个文本的相似性就可以通过它们相交项集对于这两个文本主题的贡献程度来衡量。因此，对于文本 d_i 和 d_j，根据以下情况可以判断它们之间的相似程度。

（1）如相交特征集 $T_{di} \cap T_{dj} = \Phi$，则表明文本 d_i 与文本 d_j 在主题上没有共性，相似度为零。

（2）如相交特征集 $T_{di} \cap T_{dj} \neq \Phi$，则表明这两个文本在主题上有相似性，然后用以下公式计算它们之间的相似程度：

$$sim(T_{di} \cap T_{dj}) = \left(\frac{Weight(T_{di} \cap T_{dj})}{Weight(T_{di})}, \frac{Weight(T_{di} \cap T_{dj})}{Weight(T_{dj})} \right)$$

计算公式说明，如果两个文本之间存在主题上的共性，那么它们的相交特征集中应该有对于这两个文本主题贡献较大的若干公共集合元素。也就是说，$Weight(T_{di} \cap T_{dj})$ 越大，则表明文本 d_i 和类 d_j 的相似程度也越大。

3. 划分准则

基于划分的聚类方法中，给定一个文本集 $D = \{ d_1, d_2, ..., d_i, ..., d_n \}$ 和一个期望的聚类数量 K，还有一个相似度度量方法和一个划分准则。本聚类方法中，就按照以上相似度公式计算文本与文本以及文本与不同簇之间的相似度，并根据相似度大小把文本划分到与它最相似的那个簇中。相关术语定义如下：

（1）$D = \{d_1, d_2, ..., d_i, ..., d_n\}$：$D$ 为文本集，d_i 为 D 中第 i 个文本。

（2）$C=\{C_1,C_2,...,C_i,...,C_m\}$：$C$ 为聚类特征集，其中 C_i 为第 i 个类特征集。

（3）$CD=\{CD_1,CD_2,...,CD_i,...,CD_m\}$：$CD$ 为聚类结果文本集，其中 CD_i 为第 i 个类的文本集。

（4）$FD=Feature\_of(D)$：特征提取的形式化定义，FD 为特征集合。

4. 多策略聚类算法

本聚类算法是基于 Step-by-Stet 的划分聚类，但根据给定的不同聚类参数采取不同策略。

（1）自然聚类：对于文本集 D ，开始时只有一个类，即 $C=\{C_1\}$，同时将第一个文本 d_1 划分到第一个类 C_1 中。然后从 D 中提取下一个文本 d_2 并计算与 C_1 的相似度，如相似度大于给定阈值，则将 d_2 划分到 C_1 中，否则生成新的类并把它划分到新的类 C_2 中。以此类推，依次计算 d_i 与 C 中每一个类的相似度，一旦发现相似度大于给定阈值的类，就将 d_i 划分到该类中，否则形成新的类，并将 d_i 划分到新的类中，直到全部被划分完为止。

（2）K 聚类：是指根据期望的聚类数量 K，将文本集划分成 K 个类。首先进行自然划分，然后将文本个数最多的 K 个类作为中心再进行合并。

（3）K&M 聚类：是指根据期望的聚类数量 K 和每一类最大文本个数 M，将文本集划分成 K 个类。首先进行自然划分，然后将文本个数最多的 K 个类作为中心，在每一类文本个数不超过给定 M 的前提下再进行合并。

（3）K&M 聚类：是指根据期望的聚类数量 K 和每一类最大文本个数 M，将文本集划分成 K 个类。首先进行自然划分，然后将文本个数最多的 K 个类作为中心，在每一类文本个数不超过给定 M 的前提下再进行合并。

本算法在自建数据集上与经典聚类算法 K-means 对比实验结果如表 2-1 所示。从 F 值来看，GAAC 效果最好，然后是本算法 K&M 聚类和 K 聚类，K-means 效率最差。从聚类时间效率对比中可以看出，聚类速度最快的是本算法 K&M 聚类和 K 聚类，能较好的满足大数据文本挖掘中的聚类需求。

表 2-1 聚类算法对比实验结果

指标 / 聚类算法	P	R	F-measure	聚类速度
GAAC	91.6%	90.5%	91.0%	1.7k/s
K-means	84.4%	28.2%	42.3%	19.4k/s
K 聚类	59.2%	58.2%	58.7%	95.3k/s
K&M 聚类	63.1%	63.1%	63.1%	112.2 k/s

2.1.3 维吾尔文自动摘要

与以上聚类类似的方法，评价和选取语义串作为关键词，研究了一种抽取式单文档自动摘要方法。关键句抽取时，用以下公式计算候选句权重。

$$W(s,d)=\sum_{i=1}^{n(n<=20)} W(k_i,s)$$

其中，$W(s,d)$是文本 d 中句子 s 的权重，$W(k_i,s)$是句子 s 中第 i 个关键词（语义串）k_i 的权重，n 为句子中关键词个数。对于冗余句排重问题，将候选句 A 和 B 分别看成关键词集 U_A 和 U_B，然后用 $U_A \cap U_B$ 对于 U_A 和 U_B 的比例来衡量句子 A 和 B 之间的相似度。如将句子之间的信息冗余度量化为$|U_A \cap U_B|$，则句子 A 和 B 相对于 A 的冗余度表示为：

$$Overlap\_A=\frac{U_A \cap U_B}{U_A}$$

最后，根据句子对齐信息获取每一个文摘句在原文中的位置信息，并按原文中的顺序输出。维吾尔文摘要抽取流程如图2-1所示。

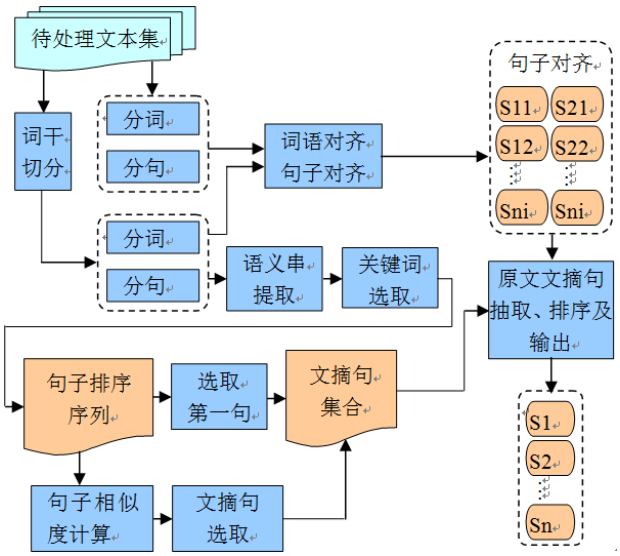


图 2-1 摘要抽取流程

对于 6867 篇维吾尔文文本（文本大小≤9KB）进行摘要抽取实验结果表明，摘要抽取时间在 5ms 以内，摘要质量也基本接近人工摘要质量。研发了一个抽取式单文档自动摘要系统，对于短文本和长文本同样有效，为维吾尔文句子级的文本处理提供支撑。

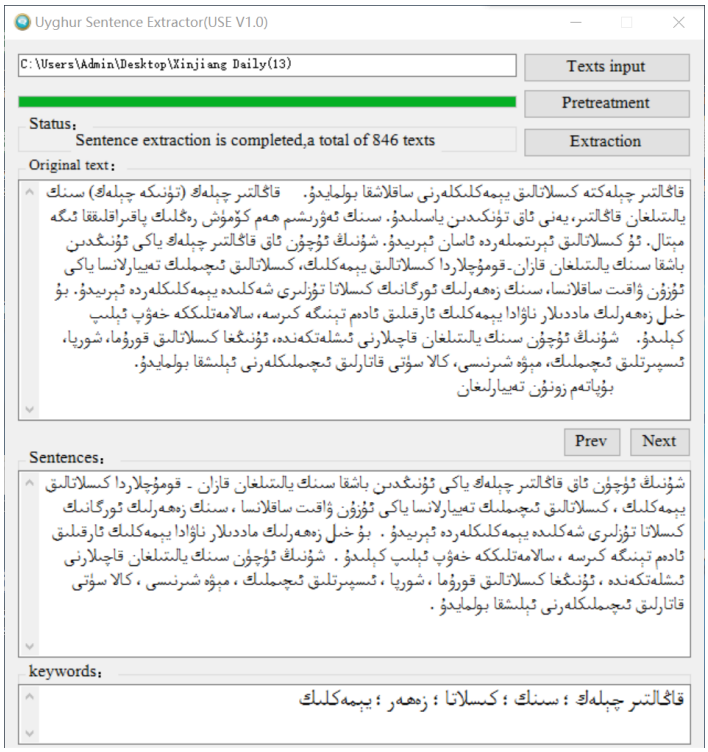


图 2-2 对于短文本的关键词及摘要抽取效果

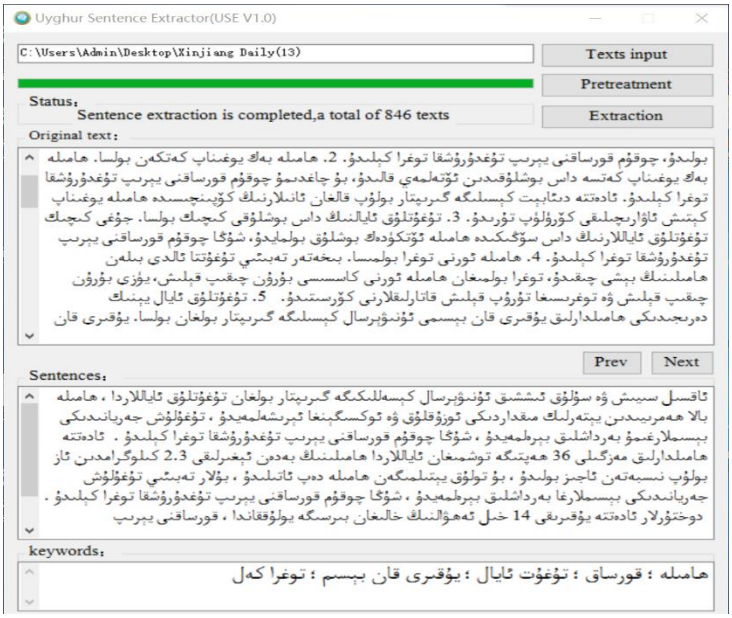


图 2-3 对于长文本的关键词及摘要抽取效果

近年来，以 BERT 为代表的大型预训练语言模型向 NLP 新范式迈上大步，已开启了语言文字信息智能处理的新篇章，各项自然语言处理任务也取得了更好的成绩。这些大模型的出现，为汉语、英语等大语种 NLP 任务提供了迁移学习的模型和解决思路，对于那些通关

基准测试的难题和企业应用，已成为了当前最好的选择。但可惜的是，这些语言模型在维吾尔文等资源贫乏的小语种领域的预训练仍在等待“数据量”的到来。

因此，通过多种途径构建模型训练用高质量维吾尔语语料库，研究模型训练中适合于维吾尔语的切分方法，训练出多种（如 BERT 等）预训练模型，从而做到在 NLP 多个任务上能够与汉语同步发展，是维吾尔语文本信息智能处理要努力发展的方向。

2.1.4 维吾尔语本体构建

维吾尔语的本体研究才开始，况且构建本体存在着方式多样、领域区分等现象使得本体共享以及重用受到了限制。为此打造本体构建规范，是实现本体顺利构建和大规模发展的重要前提，进而保证知识组织上本体能够发挥最大优势，给知识的分析、知识的检索、知识的存储创造有利条件。

维吾尔语的本体构建经历了最初的人工构建到半自动构建，最后实现自动构建，下面分别介绍每一个过程中的成果、存在的问题及未来的研究方向。

1. 维吾尔语领域本体人工构建

领域本体的构建基于人工并辅以领域专家的帮助，技术上比较简单，在领域本体中，包括该领域的所有概念。一般情况，将领域的名称作为该本体的顶级概念,并且在本体中的概念以概念层次的方式组织。这样，本体中包括了概念之间的一般层次关系。经研究本体构建所有方法和理论，选择本体构建的改进的七步法与维吾尔语的特点相结合，使用 Protégé 5.0 首次人工构建关于信息科学和数学领域的维吾尔语领域本体知识库。具体构建流程如下图 2-4 所示：

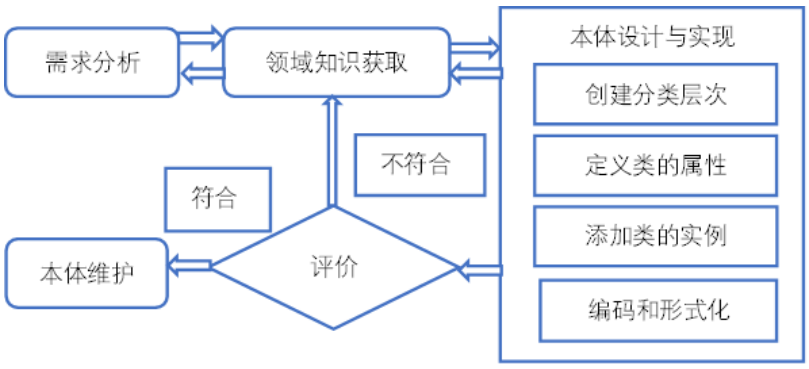


图 2-4 维吾尔语领域本体构建流程

本次构建的两个领域本体为信息科学（ISO）和数学（MO）领域，ISO 包含 14 个大类，899 个子类，100 个实例，60 个属性；MO 包含 6 个大类，92 个子类，885 个实例，6 个属性。在 Prot é g é 5.0 中的数学领域所有概念及关系可视化截图如下图 2-5 所示。

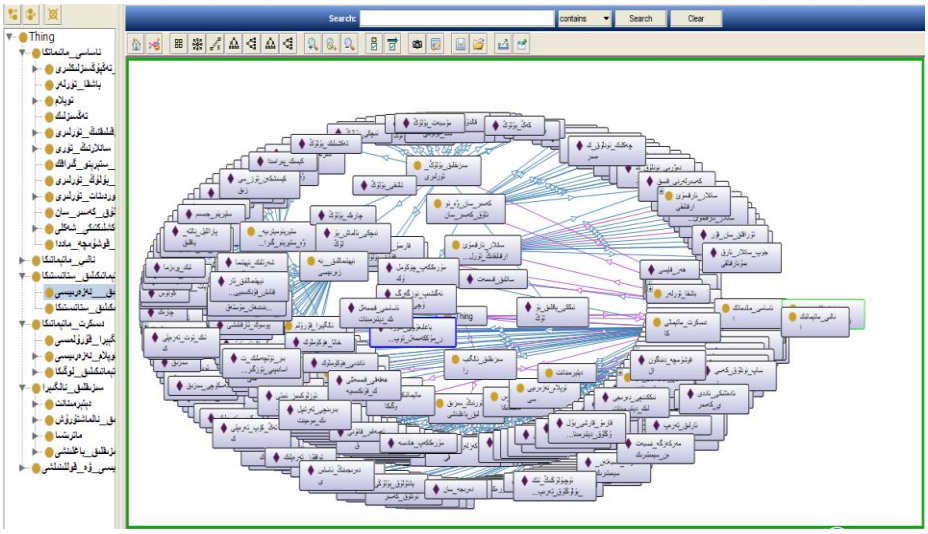


图 2-5 数学领域本体所有概念及关系

由于本体构建方法没有统一的和规范的标准，不同的专家有不同的本体论模型，即使在同一领域，他们也可以在不同的视角或层次上描述该领域的知识。因此构建本体时存在一些问题，如，需求不足，本体异构性，和在重用和共享知识时，需要在不同语言本体中进行本体集成等。

2. 维吾尔语统用本体库的半自动构建

由于维吾尔语中缺乏类似于 WordNet、Howet 之类的结构化的词汇知识库，影响了领域本体构建中现有知识库的重用率。因此开发了一种名为维吾尔语 WordNet（简称 UWN）的词汇本体工具，包含维吾尔语中典型的词汇和语义关系，包括同义关系，上位/上位关系，反义关系等。词性考虑类似与 WordNet，即名词、动词、形容词和副词四种词性，结果与 WordNet 的概念之间的匹配准确性进行了统计。

构建 UWN 的具体流程包括：（1）英语词汇的收集；（2）跨语言翻译；（3）关键词匹配；（4）获取维吾尔语；（5）验证和索引；（6）UWN 中关系的确定。我们从英语 WordNet 中提取所有关键词及其相关术语，事实上，WordNet 中的关很多种，但到目前为止，我们只考虑同义词，上位词/下位词和反义词。扩展到其他关系是今后的研究方向。

由于 WordNet 中的关键词汇量非常大，即 155,287 个，那么开发一个系统来自动映射这些语言的所有相关关键词并非都可以正确映射，因为单词不是一对一对应的。因此手动检查并映射未正确映射的单词，否则对于错误映射单词无法纠正。在映射时存在模糊映射的特殊情况，即，多个单词被映射到多个其他不同的单词。为了解决此问题提出了正反匹配策略来处理，从而解决词汇歧义问题，由图 2-6 所示。

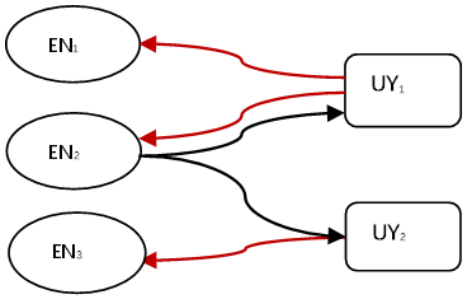


图 2-6 词义映射正反匹配策略

图 2-7 UWN 用户查询界面

由图 2-7 所示是 UWN 用户查询界面。系统完成后，对实验研究评估系统的性能，选择了一个由 5 名研究人员组成的团队来评估。虽然本工具无法完全提供 WordNet 中提供的所有功能，但可以在最短的搜索时间范围内稳健地提供所输入关键字的同义词。

表 2-2 UWN 和 WordNet 中关键词匹配率

	关键词	同义词集	名词	动词	形容词	副词
WordNet	155287	117659	117798	11529	21479	4481
UWN	133063	99101	96947	11214	20466	4436
匹配率	85.68%	84.23%	82.29%	97.27%	95.28%	98.99%
平准						90.67%

由表 2-2 可见，本工具对所有实现的功能都具有 90%以上的准确性，这表明本工具中提供的映射基本上几乎与 WordNet 相匹配。但由于英语和维吾尔语中名词的使用不同，名词的匹配准确度较低。在英语中，大多数动词都可以在句子中使用名词形式表示，但在维吾尔语中使用名词有严格的语法要求。所以 UWN 名词的数量相对少于 WordNet 名词。

3. 维吾尔语领域本体库的自动构建

本体自动构建包括两个步骤，概念自动获取和关系自动获取。

（1）维吾尔语领域本体概念的自动获取

随着信息技术的发展和新知识的出现，本体无法满足人们不断变化的需求。此外，增长和维护本体将是一个具有挑战性的问题，本体工程师必须使用各种领域特定的知识来维护本体的更新并坚持现有本体的总体结构。为了机器自动理解纯文本并从中提取所需的知识，应该通过自然语言处理技术进行预处理，然后通过统计和机器学习等技术获取相关知识，然而纯文本缺乏一定的数据结构。因此提出了通过尝试词频统计 TF-IDF 的基础上整合互信息（mi）、置信度(Cd)和邻接熵（Ea）来提取维吾尔语领域本体的概念，以便有效提高此方面本体概念提取的自动化程度和正确率。

互信息（Mutual Information）：指衡量信息相关性的方法。应用到文本中的信息时，衡量两个词汇的相关性.

$$mi(A,B) = log2\frac{P(A,B)}{P(A)P(B)}$$

其中, P(A, B)为词对(A, B)在大规模语料库中出现的概率, P(A)为单词 A 出现的概率, P(B)为单词 B 出现的概率。为方便分析，而假定其在语料库中出现词的频率分别为count(A)、count(B)、count(A, B)，n 是语料库中的词频总数。

$$P(A,B) = \frac{count(A,B)}{n}, P(A) = \frac{count(A)}{n}, P(B) = \frac{count(B)}{n}$$

置信度（Confidence Degree）：是过滤数据的基础，选择置信度最高的词作为自动匹配结果。置信度是出现单词关联（wi-1 wi）的上位（前面）wi-1 的情况下 wi 出现的条件概率。它定义如下：

$$\mathrm{Cd}(w_{i-1}\ w_i) = \mathrm{P}(w_{i-1}/w_i) = \frac{(w_{i-1} \cup w_i).\mathrm{count}}{w_{i-1}.\mathrm{count}}$$

邻接熵（Entropy of Adjacency）：先确定出单词串的左右相邻熵，如果对比结果表明此数值超过阈值，则相应的单词串被认为是独立语

言单元，并且考虑单词串。本文将上述思想引入维吾尔语研究，发现它适合于文本的研究需要，通过“连接”和“破坏”基于熵的词来判断问题。Ea(A,B)取值越大，词对 A 和 B 的语言环境变化越灵活多样，其内部结合越紧密； Ea(A,B)取值越小，A 和 B 的独立性越弱，在一些情况下可能是偶然组合而形成的，因而若 Ea(A, B)>T 时,这两个单词间位置一般为“连”，相反情况下判断为“断”。

$$Ea = -\sum_{i=1}^{c} \frac{n_i}{m} \log \frac{n_i}{m}$$

通过计算语料库中每组相邻词的互信息、置信度和邻接熵来获得二元关系参数数据集。每个参数都有自己的优点和缺点。因此，为了提高领域词汇的准确性，我们考虑组合使用三个参数。根据阈值计算相邻单词的最终权重，然后在分割后提取领域词汇。切分完成后，对分段文本进行词频统计（TF-IDF 算法），最后过滤权重较高的单词作为提取的领域概念。表 2-3 所示仅针对旅游领域的领域概念提取的统计情况，结果表明在综合使用三个参数（Mi, Cd, Ea）之后再加上 TF-IDF 算法来提取的概念的数量已经增加。

表 2-3 三个方法的对比

	提取数量	**正确数量**	**准确率(%)**	**召回率(%)**
Mi	310	212	68.4%	62.9%
Cd	193	105	54.8%	31.2%
Ea	295	201	68.2%	59.6%
Mf **+ TF-IDF**	307	255	83.1%	75.7%

然而 UWN 也存在一些不足之处：

①除了同义词，上位词，下位词和反义词之外，还需要考虑所有其它关系，并需要研究适合维吾尔语的新关系。

②将 UWN 商业化以供行业使用。

③需要进一步扩展 UWN 的各个性能模块，以供支持国内类似的属于

同一个语系的少数民族语言，如哈萨克、柯尔克孜、塔塔尔等。

（2）维吾尔语领域等级关系的自动过获取

概念间关系主要包含等级关系和非等级关系两类。其中等级关系是指领域概念相关的隶属关系、包含关系、位级关系等，可对相关的概念体系进行描述。进一步分析可知，等级关系是指两个概念间的各方面关系之和，如果一个概念包含的对象在另一个概念之内，则可认为二者存在一定的等级关系。举例来说概念“旅游”和概念“旅游景点”存在这种关系。非等级关系也称相关关系，可以将其看作为一种联想关系，即基于人类的知识积累而建立的，相应的范围很广。比如：两个概念的意思相反，则二者就表现出反义关系。此外概念“老师”和概念“学生”之间就存在“教”或“被教”的语义关系，也是一种非等级关系。概念间的这些关系可在相应的领域文档中进行提取，也可以通过相关的语法特征进行挖掘。

我们主要从等级关系的获取方法上进行深入的讨论。这种等级关系从本质上看，就是领域本体中不同对象相关的连接关系，可看作为领域概念对应的分类关系。通常，可通过树状图来表示层次聚类流程。其可反映出对象是如何一步步分组的。图 2-8 中显示五个对象｛A、B、C、D、E｝的聚类过程，其中 A、B 根据相似度合并为一个类{A、B}，D、E 也根据相似度合并构成另一个类{D、E}，这个类再和 C 合并后再和{A、B}类合并最终整个聚类成一个大类。

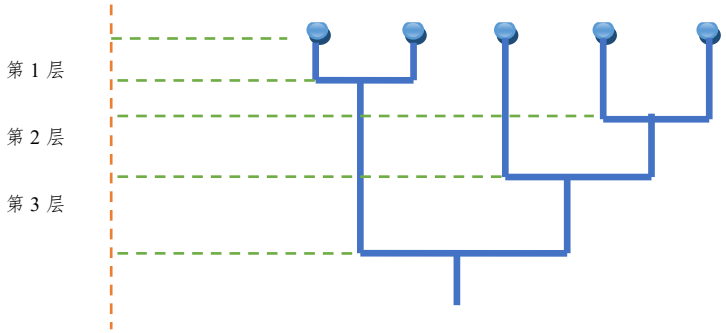


图 2-8 数据对象｛A,B,C,D,E｝层次聚类树状图

我们在计算层次聚类时为了提高准确度充分利用领域叙词表的层次结构为领域本体的主题层次结构，在此基础上利用层次聚类法的相似度算法，综合计算概念间的欧氏距离和切比雪夫距离，一直重复直到最后合并成一个类为止，最后实现文本中的领域词汇的层次划分。两个n维向量a与 b的欧氏距离和切比雪夫距离计算公式如下所示。

$$d_{12} = \sqrt{\sum_{k=1}^{n}(x_1k - x_2k)^2}$$

$$\mathrm{d_{12}} = \lim_{\mathrm{k\to\infty}}\left(\sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{x_{1i}\text{-}x_{2i}}|^{\mathrm{k}}\right)^{1/\mathrm{k}}$$

在进行层次聚类过程中选择了自底向上的方法，也就是先将每个对象作为一个单列的原子簇，然后依据一定的方法不断的进行合并而得到规模大的簇，不断的进行重复操作，直到全部的类都在一个簇中（层次的最上层），或者达到一个，或者相应的终止条件满足为止。我们利用混合层次聚类法提取的等级关系结果表2-4所示。根据此方面的经验可知，这种方法确定出的等级关系可满足本体等级关系的要求，但还需要完善系统性能和层次精度。

然而表2-4中数据可见最终层数和聚类的节点数差别比较大，存在不足。经分析原因有以下几点：

1）文本语料的内容没有多样性和广泛性。比如，房产类的文本内容基本都是买房、卖房、求组和出租之类的信息。因此内容的局限性和重复性，聚类提取的节点数比总文本提取的概念数少很多，分层数只有 4。明星类的层数 9，聚类概念数相对最多是 121，原因是原始文本里包含国内外很多明星不同信息，人数比较多，所以提取的不同概念也比较多。

2）文本内容存在的一些拼写错误和不规范性。由于文本收集是由维吾尔语网站上的广告和新闻等途径下载的，而且维吾尔语文是黏着性语言书写时容易出现拼写错误。

3）聚类中心点坐标计算的精度不是很高。本文设置的向量维度

不高导致 word2vec 训练出的关键词的向量精度不高。

4）层次聚类法不具有很好的可伸缩性，因为合并或分裂的决定需要检查和估算大量的对象或簇。

5）层次聚类方法很容易操作，不过在实际的计算过程中经常出现不容易确定合并或分裂点的问题，而这种点选择对其后的操作结果会产生直接的影响，在其后的操作过程中需要对新生成的簇处理。

表 2-4　不同类用层次聚类生成的层次

领域	层数	聚类概念	总 概 念	占总数
旅游	7	51	500	10.2%
健康	9	99	500	19.8%
教育	8	91	500	18.2%
体育	10	55	500	11%
房产	4	43	500	8.6%
明星	12	121	500	24.2%
交通	9	81	500	16.2%
汽车	8	52	500	10.4%
文艺	9	40	500	8%
招聘求职	6	41	500	8.2%
手机	7	65	500	13%
经济	9	111	500	22.2%
计算机	7	81	500	16.2%

在一个完整的领域本体构建中非等级关系的研究是必不可少的。由于文本语料的稀疏和时间的有限等原因没有对非等级关系提取作出探讨，今后的研究将着力于非等级关系及其提取方法，以及扩展到知识图谱的构建及应用上。

2.2 蒙古语智能信息处理

蒙古语智能信息处理工作是内蒙古自治区信息化建设的重要内容，对于宣传贯彻党和国家民族政策法规，铸牢中华民族共同体意识，

打造祖国北疆亮丽风景线具有重要意义。近年来，内蒙古自治区党委、政府高度重视蒙古语言文字信息化工作，制定出台《内蒙古自治区人民政府关于加快推进蒙古语言文字信息化建设的意见》（内政发〔2012〕106 号）《内蒙古自治区蒙古语言文字信息化建设中长期规划（2014-2020 年）》（内政办发[2014]37 号），促进蒙古语言文字信息化基础研究、人才培养、技术研发、资源建设、推广应用等各领域全面协调发展，取得了积极成效。形成了政策扶持、规划引导、项目带动，部门配合、社会参与，合力推进蒙古语言文字信息化建设的良好局面。

2.2.1 资源建设

1. 蒙古语言文字资源建设成果丰硕

建设完善了蒙古语粗加工语料库、平行语料库、语义信息词典、熟语知识库、语句法结构知识库、方言口语语料库、口语韵律标注库、语音数据库、蒙汉文可比语料库等蒙古语言文字资源库，建设了统一编码、统一标准规范的蒙古语资源基础库，为蒙古语言文字信息化基础研究和应用开发提供了可靠的语料、准确的标准；形成了涵盖蒙古语婴幼儿智能早教、蒙古语授课中小学信息化教学、蒙古文多语种经济管理教学、蒙古文词典百科全书等综合工具书、蒙古语言文字数字化学习、蒙古文 MOOC 教学等丰富的教学资源库，进一步扩大现代远程教育覆盖范围；开展了多行业、多领域、多维度的蒙古文图书、音视频、民俗资料、历史报纸、历史档案、古籍文献等相关资源的数字化工作，形成了统一规范和标准的蒙古语言数字资源库。

2. 蒙古文古籍文献数字化加工成果显著

2019 年 7 月 16 日，习近平总书记考察内蒙古大学时强调，要加强对蒙古文古籍的搜集、整理、保护，挖掘弘扬蕴含其中的民族团结进步思想内涵，激励各族人民共同团结奋斗、共同繁荣发展。为贯彻落实习近平总书记的指示精神，内蒙古大学联合内蒙古社会科学院 2020 年启动了“蒙古文古籍文献数字化工程”项目。2021 年 10 月项

目初步设计方案通过了自治区发改委审批，获批自治区政府重点基础建设项目。目前编制完成了《蒙古文古籍文献数字化标准及规范》，并已完成 400 多部蒙古文古籍文献的数字化、12 部珍贵古籍的全文数字化。搭建了蒙古文古籍文献文字识别和知识库原型系统，完成了蒙古文古籍文献数字化及智能分析平台的设计工作，并正在进行系统研发。经过项目建设，将建成一个存储数字化蒙古文古籍文献最多、检索分析智能化程度最高、相关教学和科研人员最依赖的国际一流的蒙古文古籍文献数字化中心，为蒙古文古籍的抢救、保护、传承、开发和利用提供重要支撑。为我国民族文化建设、国内外蒙古学与民族学教学科研和自治区文化经济建设服务。

相关机构还开展了蒙古文古籍文献、档案资料数据库建设，开发电子书应用产品，采取复制、缩微、电子数字化等技术手段，推进蒙古文古籍文献、档案资料的数字化保护与共建共享。主要工作包括内蒙古图书馆蒙古语视频资源建设、内蒙古图书馆蒙古文公益服务管理平台建设、中古蒙古语多文种文献语料库建立与数字化平台搭建、卫拉特研究相关蒙古文文献数字化管理与共享平台研发、基于大数据的蒙古马文献数据资源开发应用、蒙古语古文献专业字体数据整理与制作工程、全媒体时代内蒙古少数民族古籍数字化传承和弘扬工程、蒙古文历史报纸数字化及内容资源公共服务平台建设、内蒙古大学民族博物馆蒙古语数字化项目等。进一步扩大了国家和自治区少数民族语言文字出版物数字化规模，初步实现了面向公众服务的蒙古文文献检索、聚合与导航、信息推送和信息发布的公益平台。

3. 蒙古语言文字数字资源共享平台

为充分满足各族干部群众对蒙古文软件应用、公共服务和信息需求日益增加的实际需求，认真组织实施蒙古语言文字数字资源建设与共享工程项目。实施蒙古语言文字数字资源建设与共享工程项目是《内蒙古自治区人民政府关于加快推进蒙古语言文字信息化建设的

意见》《内蒙古自治区蒙古语言文字信息化建设中长期规划（2014—2020年）》明确的重点任务。内蒙古自治区发展和改革委员会于2014年9月批准立项，2015年11月批复同意《蒙古语言文字数字资源建设与共享工程项目初步设计方案和投资概算》。2016年6月至11月，通过内蒙古自治区政府采购中心完成公开招标，被评为2016年度全国政府采购精品项目。2017年至2019年，按照统筹规划、标准先行、分步实施的原则，分阶段完成了项目建设任务，2020年6月上线试运行，10月通过内蒙古自治区发展和改革委员会验收。该项目以蒙古语言文字数字资源开发整合为基础，以蒙古语言文字信息技术规范化标准化为支撑，以蒙古语言文字数字资源推广应用为目的，开发建设标准规范、公共基础支撑系统，应用服务系统和数字资源，建成蒙古语言文字数字资源共建共享综合性公益服务平台，完成了包括马克思主义、毛泽东思想、邓小平理论、“三个代表”重要思想、科学发展观、习近平新时代中国特色社会主义思想以及党和国家法律法规、文化、教育、科技等内容的数字资源，共计文本资源1,034,623千字、图片473,185张、音频120,027分钟、视频186,098分钟、课件2,524个、动漫6,385分钟，资源数据总量49.87TB。上线运行后，经两年的建设，现资源容量已扩容至近100TB，点击量达到100万人次。建成蒙古语言文字数字资源共享平台，为各族群众提供方便快捷、优质高效的数字化、信息化公共服务；实现了蒙古文工具书在线服务，为广大蒙古语文工作者和区内外各族群众提供优质便捷的数字化在线服务。内蒙古自治区民族事务委员会将推广应用工作作为各盟市蒙古语言文字信息化工作重点来部署，广泛调动各地区积极性和主动性，推动蒙古语言文字信息化推广应用工作，为基层群众生产生活提供方便、快捷的政务服务、资源服务和信息服务积极创造有利条件。

2.2.2 蒙古语语音识别技术

蒙古语从书写形式上可分为两种：在中国境内使用回鹘体书写，称为传统蒙古文；在蒙古、俄罗斯等地使用西里尔字母拼写，称为西里尔蒙古文，二者语义相同而文字表现形式不同。

与汉语、英语以及阿拉伯语等使用人口多且普及率较高的语言相比，蒙古语这种小语种的语音识别技术的研究在国内外均起步较晚，目前获得的关注也较少。在国外，蒙古语语音识别研究几乎只关注于西里尔蒙古文的识别问题。2009 年日本的情报通信研究机构（National Institute of Information and Communications Technology，NICT）[1][2]针对西里尔蒙古文提出使用相似词分类多类 N-gram 模型对语言模型进行研究。在已有的包含多种语言形式的语音自动翻译系统中追加了蒙古语模块，实现了蒙古语识别系统，该方法相比于使用传统的 N-gram 语言模型，单词错误率相对下降了 5.5%。近年来，Mozilla 的 common voice 项目包含了西里尔蒙古文的数据采集计划，目前已收集语料 18 小时，其中 12 小时经过标注，以此为基础，不少新的语音识别算法被应用于西里尔蒙古文的语音识别中，据 papers with code 网站的排行榜，目前最优系统的识别错误率尚高达 34.64%（截止 2022 年 6 月）。

在国内的蒙古语语音识别研究主要集中于传统蒙古文，与经过标准化的西里尔蒙古文不同，传统蒙古文的文字拼写在其漫长的发展历程中形成了一些独有的特色：传统蒙古语单词存在多音字，一些单词在不同的上下文中会有不同的发音；反过来，有些不相同的发音也可能对应着相同的单词；虽然传统蒙古语是一种拼音文字，但发音与文字拼写并不严格对应，在发音中经常出现元音和辅音的增加、脱落及变换等问题。内蒙古大学计算机学院高光来教授带领着科研团队于 2003 年开始对蒙古语语音识别相关工作展开研究。随后，包世恩[3]、毕为格[4]以及哈斯其劳图[5]在蒙古语语音语料库的构建、语料标注以

及发音字典构建作出了一定的贡献，并以音节、单音素和三音素等为基本单位构建声学模型，使用 N-gram 统计语言模型，搭建了简单的蒙古语连续语音识别系统。之后，飞龙[6-8]在研究蒙古语大词汇量语音识别相关工作中，提出基于分割的方法将蒙古文单词切分成词干及后缀形式并分别作为基本识别单元。通过对声学模型以及语言模型进行重建，相比原始建模单元构建的系统取得了较高的识别性能提升，为之后的使用词干及后缀作为基本建模单位的蒙古语语音识别研究奠定了基础。最近，随着深度学习方法在模式识别相关任务中的兴起，语音识别的研究进程得到了飞速发展。张晖[9]等人在蒙古语语音识别研究中首次引入了 DNN 神经网络，识别性能相比传统方法提升了将近一倍。张红伟[10]等人将 TDNN、CNN 和 LSTM 等应用于蒙古语语音识别声学模型，识别性能进一步得到提升并达到了实用效果。王勇和等[10]针对蒙古文的书面形式转口语形式时元音和辅音的增加、脱落及变换等问题，提出了基于注意力机制的编码器解码器模型结构的蒙古文字母到音素转换方法。更进一步，为了解决注意力学习的非单调对齐问题，基于字母-音素之间是单调对齐关系创新性的提出了对齐学习损失函数，提升了转换模型的稳定性。

2021 年，西北民族大学、清华大学联合发布了目前最大的公开传统蒙古语语音识别语料库[12]，其中包含传统蒙古语数据 170 小时，使用 GMM-HMM、TDNN-HMM 的最优识别模型的 WER 分别为 62.2%和 52.5%，CER 分别为 27.1%和 19.7%，使用最新的端到端模型（Transformer）的性能稍好，WER 和 CER 分别为 48.0%和 11.6%。2022 年，上海交大钱彦旻团队[13]也研究了蒙古语的语音识别问题，在一个较小的数据集上的 CER 为 53.7%。目前在小规模数据集上的蒙古语语音识别工作均无法达到实用水平。因此，语料库的建设成为了制约蒙古语语音识别研究的关键问题，内蒙古大学蒙古文信息处理技术重点实验室计划在 2022 年公开 345 小时的蒙古语语料，并将持

续为蒙古语语音识别的研究工作贡献新的资源。

2.2.3 蒙古文文字识别技术

蒙古文的信息处理工作在少数民族语言文字信息处理领域中起步较早，在 80 年代就已实现计算机上的蒙古文信息处理系统，为内蒙古自治区推广计算机蒙古文信息处理和蒙古文数字化创造了良好的条件。随着移动互联网和人工智能技术的发展，蒙古文文字识别研究成为了蒙古文信息处理的主要研究方向。

蒙古语属于黏着语，即单词的主干由所有字母在垂直方向上粘合在一起形成，且字母根据它们在单词中的位置呈现出不同的样式。蒙古语单词的构词和词形变化是通过将不同的后缀连接到词根或词干来实现的，这使得蒙古语的词汇量相当庞大，经常使用的词汇量约为 100 万个。此外，蒙古语与阿拉伯语、英语和其他拉丁语言有很大的不同，它的书写顺序是从上到下垂直的，文字列的阅读顺序是从左到右。根据蒙古语的这些特点，早期针对其的文字识别工作均是基于切分的，而随着识别技术的不断发展，整词识别成为了研究主流。

所谓基于切分的识别方法，需要先对输入的单词图像进行切分，之后对切分后的部分进行特征提取，进而识别每个切片所对应的字元。在 2000 年，内蒙古大学蒙古文信息处理课题组高光来教授等人首次针对印刷体蒙古文进行了基于切分方法的研究[14-17]，通过特征设计、字元切分、特征提取以及特征匹配多个步骤完成识别工作，在常用的蒙古文印刷字体（如白体）中实现了 90%的单词识别准确率。之后，高光来教授针对木刻版蒙古文古籍的手写体单词进行了基于切分的识别研究[18]，通过决策树以及 BP 神经网络技术，分三个阶段对手写体蒙古语单词进行识别，在提高识别准确率的同时填补了该领域研究的空白。苏向东等人在此基础上针对木刻版蒙古文古籍的文字特点进行了分析[19]，提出了蒙古语单词的主干线定位和切分方法，并在切分

的基础上为构成最终的识别结果选定了最小字元集。接下来，苏向东等人在[20]中改进了[19]的切分方法，通过单词外轮廓显著轮廓点检测、基于 Logistic 回归模型的基线定位以及基于启发式规则的候选分段路径生成三个步骤对古籍单词图像进行切分，将字符层面的切分正确率提升到了 89.84%。为了提高蒙古文古籍的识别性能，苏向东等人提出了基于知识的多策略融合古籍识别系统[21][22]，从欠分段与过分段片段的识别、字元分组以及基线信息合并三个角度共同作用，提升单词识别的准确率。范道尔吉等人则提出了基于隐马尔可夫模型和深度神经网络混合结构的大词汇量蒙古文脱机手写识别系统[23]，在手写体蒙古语单词的识别性能最高可达到 97.61%。

伴随着深度学习技术的蓬勃发展，考虑到基于切分的识别方法步骤繁琐以及切分效果不佳等问题，针对蒙古语单词的整词识别方法开始成为主流的研究方向。整词识别的方法使用深度神经网络提取输入的文本图像的特征，通常使用卷积神经网络进行特征提取，再使用循环神经网络进行序列建模，最后利用连接时序分类对输出的序列进行处理，得到最终的识别结果。张晖等人提出了一种针对传统蒙古文单词的整词识别方法[24]，将 OCR 任务形式转化为序列到序列的映射问题，即将图像帧序列转化为字符序列，在实验中证明了该方法的有效性。Wang 等人提出了基于端到端模型的整词识别方法[25]，同时在集外词问题的解决上也有不俗表现。魏宏喜等人针对木刻版蒙古文古籍整词识别提出了一种基于卷积神经网络的的识别方法[26]，同时为了解决数据集上样本分布不均衡问题，使用 SMOTE 技术生成样本，也为解决集外词问题提供了新思路。Kang 等人同样针对木刻版蒙古文古籍识别提出了一种具有注意力机制的序列到序列模型[27]，使用深度神经网络和双向长短期记忆网络编码输入的文字图像，通过注意力模块提取关键图像帧特征，再通过解码器解码为识别的字符序列。在脱机手写体蒙古文识别领域，魏宏喜等人提出了一种端到端的带有注意力

机制的体系结构来执行从原序列生成目标序列的任务[28]，从而提高了脱机手写识别的性能。这些方法也应用在大词汇量脱机手写体蒙古文整词识别和联机手写体蒙古文识别工作中[29][30]。为了进一步提升脱机手写体蒙古文识别的准确率，范道尔吉提出了一种基于序列到序列模型和基于子词的语言模型的识别系统[31]，从单词图像的预处理，图像到字素序列的映射，以及基于子词的语言模型(LM)解码三个方面提高了脱机蒙古文手写体识别的准确率。针对蒙古文手写单词具有序列数据特点以及变形严重问题，范道尔吉提出了隐马尔可夫模型与深度神经网络相结合的混合识别方法[32]，将每个脱机蒙古文手写单词都看作沿书写方向的一个一维随机序列，通过隐马尔可夫模型描述该序列的生成过程，深度神经网络描述序列的概率分布，将语音识别相关方法成功移植到蒙古文脱机手写识别任务中，取得了较好效果。魏宏喜等人则将多任务的思想融入到了识别任务中[33]，所提出模型可以同时完成蒙古文字形切分和蒙古文单词识别任务，不仅提高了单词识别的性能，而且提高了字符分割的准确性。Cui 等人面向不规则蒙古文的识别问题，提出了一种用于印刷体蒙古文文识别的三元注意基元网络(TAMN)[34]，使用空间变换网络来校正变形的蒙古文图像，之后采用门控递归卷积层(GRCL)结合三重注意模块对校正后的图像进行特征提取，并通过 LSTM 网络获取特征中的上下文序列信息，使得识别正确率达到 90.30%。

蒙古文单词样本的稀缺性以及单词图像存在的噪声，也在一定程度上限制了识别准确率的提升和识别研究工作的进展。对于前者，数据增广方法是一个很好的解决思路，然而普遍采用的仿射变换并不能很好地针对蒙古文单词的特点进行增广。基于此，魏宏喜等人采用对抗生成网络来进行数据增广[35]，提出了一种基于循环一致生成对抗网络的数据增广方法，使用 CycleGAN 模型学习图像到图像的转换，以生成与输入单词不同风格的样本。而 Zhang 等人则根据蒙古文单词的

特点，提出了一种蒙古文手写体单词的局部增广方法[36]，通过移动笔划的端点和笔划外控制点，使用 Bezier 曲线重构笔划，从而对笔划进行局部形变，有效地提高了增广样本的多样性。对于单词图像存在的噪声，苏向东等人提出了一种端到端的蒙古文古籍文献 OCR 预处理器[37]，采用对抗学习方式进行训练，将文档图像的预处理问题描述为一个像素到像素的问题，并用条件 GAN 来解决这个问题，使得整个过程不需要阈值计算、滤波器设计和映射函数公式，可以同时对古籍文献图像进行二值化和去噪。同时，他们还注意到了低分辨率对识别性能的影响，发现识别器对超分辨率文本图像的识别准确率明显高于未经处理的低分辨率图像，因此提出了一种基于生成对抗网络的文本图像分辨率改进方法[38]，进一步提升了识别性能。

2.2.4 蒙古语语音合成技术

语音合成解决的主要问题就是如何将文字信息转化为可听的声音信息，它涉及声学、语言学、数字信号处理、计算机科学等多个学科技术，可广泛应用于智能家居、虚拟主播、语音导航、信息播报、阅读教育、泛娱乐等领域，是人机交互的重要组成部分。蒙古语作为中国的民族语言之一，有着悠久的历史和丰厚的底蕴。它的使用人群分布在当今世界各地，包括内蒙古自治区、甘肃、西藏自治区等中国八省区，蒙古国及俄罗斯等世界不同地区。随着研究学者对文字智能信息处理研究的不断深入，蒙古语智能信息处理相关问题也受到越来越多的关注。近年来，越来越多的研究人员使用深度学习技术对蒙古语智能信息处理相关问题展开深入研究。得益于深度学习模型强大的建模能力，蒙古语语音合成的整体质量得到了显著提升。

经过研究人员的不懈努力，蒙古语语音合成技术取得了长足发展，但是相对于主流语种相关研究，蒙古语语音合成技术仍不够成熟，语音合成的前端模块和后端模块还有很多问题亟待解决。与汉语、英语

等主流语种的语音合成技术相比，蒙古语语音合成研究还有很大的探索空间，要想满足合成语音质量的实用需求，还需要更进一步的深入研究。当前蒙古语语音合成系统与真实语音相比，自然度和表现力还是明显不足，主要表现在：韵律节奏缺乏表现力，合成语音音质不够高。其中，韵律建模和声学建模能力的不足是导致这些问题的主要原因。具体来说，其原因主要源于两个方面：（1）语音合成前端模块中韵律建模效果不理想：韵律模型得到的韵律特征精度不够高，不能准确的反映出语音的停顿、快慢等信息，如果产生错误的停顿甚至会造成语义表达的错误；（2）语音合成后端模块中声学建模效果不理想：声学模型预测得到的声学参数与真实声学参数有一定差距，其合成语音因此不够清晰，错误的声学参数甚至也会造成不可预知的语义表达错误。

针对蒙古语语音合成，已有一大批学者在早期开展了大量研究。敖其尔、巩政提出了一种波形拼接的蒙古语语音合成方法[39]；高光来提出了以词为单位的波形拼接技术进一步对蒙古语语音合成方法展开研究[40]；萨其容贵将基音同步叠加法引入并建立了多样板蒙古语语音合成音库[41]；田会利提出了基于词干后缀的有限词条的蒙古语语音合成方法[42]；孟和吉雅针对蒙古语动词词干词缀，提出了另一种基于词干后缀的蒙古语语音合成方法[43]；敖敏从韵律角度出发，对蒙古语语音合成方法进行研究[44]。近几年来，统计参数语音合成方法在英语、汉语等主流语种中取得了成功应用，其合成语音的整体表现已经与真人发音非常接近。随后，基于统计参数的蒙古语语音合成技术也相继被内蒙古地区相关研究人员提出。具体的，在韵律建模方面，李婷会等提出了基于 CRF 模型的蒙古文韵律建模方法[45]；文献[46][47]等进一步利用蒙古文单词词性等符号化特征表示提升了基于 CRF 模型的蒙古文韵律模型的性能。在声学建模方面，赵建东等提出了基于 HMM 声学模型的蒙古语语音合成的方法[48][49]，该方法首先构建了蒙古语语

音语料库，结合蒙古语语言特点设计了上下文属性集以及相应于模型聚类的属性问题集，最后实现了基于 HMM 声学模型的蒙古语语音合成系统。

鉴于深度学习技术引入语音合成领域后的出色表现，一些学者开始将深度学习技术与蒙古语语音合成进行结合。针对蒙古文韵律建模，基于深度学习的蒙古文韵律建模研究还没有相关研究涉及，仍然以基于 CRF 模型的韵律模型为主。针对蒙古文声学建模，深度学习声学建模技术在蒙古语语音合成中取得了重要进展，刘瑞等人在 2017 年首次将深度学习技术引入蒙古语语音合成方法中[50]，使合成语音的音质有了明显的提升。随后李劲东、刘郅楠等人将端到端的网络结构应用在蒙古语语音合成任务中[51][52]，该方法可以直接建立文本与语音对应的关系，从而简化了建模过程。基于端到端的蒙古语语音合成技术已经能够给多家单位提供蒙古语语音合成的接口。鉴于蒙古语合成语音缺乏韵律的表现力，刘瑞等人在 2019、2020 年提出了多任务和多源知识相结合的蒙古语语音合成方法[53][54][55]。该方法基于知识蒸馏的端到端声学建模方法，采用“教师-学生”训练方式，首先训练教师模型，当教师模型训练完成后再由该模型指导学生模型的训练过程。该方法有效地降低了合成过程中跳词、漏词、重复等现象的发生。随后，刘瑞等人在 2021 年提出了通过结合自注意力机制和分类器的特征增强方法[56]，通过这种方法蒙古语合成语音的质量得到了极大的提高。除此之外，刘瑞等人通过对输入的文本及其韵律风格之间的关联进行编码[57]，进一步提高了蒙古语合成语音的自然度。这些方法的提出不仅使蒙古语合成语音越来越流畅自然，而且给后续的蒙古语语音合成技术的研究打好了基础。

这些研究方法的提出，使得蒙古语语音合成技术不断深入，合成蒙古语语音的整体表现相较传统方法也获得了显著提升，但是其合成语音的韵律表现和合成音质与真实语音相比仍然具有较大差距，其韵

律建模和声学建模部分仍有很多问题没有解决。

针对蒙古文韵律建模和声学建模，具体问题分析如下：

（1）韵律模型语义建模能力不足：现有蒙古语系统在韵律表现上较为平淡、表现力不足。具体而言，韵律建模是前端模块的重要组成部分，由其得到的韵律结构特征是合成具有丰富表现力语音的关键因素。现有蒙古文韵律建模只依赖于蒙古文单词的符号化语义特征表示和 CRF 等浅层机器学习模型，严重制约了蒙古文韵律模型的语义特征挖掘能力，更重要的是，有限的蒙古文文本资源和蒙古文独特的黏着语特性导致的数据稀疏问题，限制了蒙古文前端韵律模型的建模能力，而数据稀疏问题会进一步导致训练数据中出现大量的集外词，也给蒙古文韵律建模带来很大挑战。

（2）韵律建模缺乏相关任务的辅助：现有蒙古语语音合成系统在进行韵律建模时只使用单一模型从蒙古文单词序列及其语义特征中学习韵律结构。但是音素是蒙古文发音的基本单元，字母转音素任务中对蒙古文单词发音的音素序列进行预测，因此蒙古文单词的发音和韵律变化特征也被隐式包含其中。但是当前韵律模型并没有充分考虑两个相关联任务的深层次关系，也对韵律建模的精度产生一定影响。

（3）声学建模鲁棒性不足：端到端声学建模方法会频繁出现跳词、漏词、重复和韵律不稳定等现象。具体而言，在推理阶段，基于端到端声学模型的蒙古语语音合成系统使用上一时间步预测的声学参数进行当前时间步的参数预测，这样的解码方式导致解码误差随着时间步的推移不断积累，造成后面解码序列的不准确，生成不准确的声学参数从而导致合成语音质量下降。

（4）声学建模缺乏韵律信息显式指导：虽然基于端到端声学模型的蒙古语语音合成系统抛弃了复杂的蒙古文文本处理流程，可以直接学习蒙古文拉丁字符到蒙古语语音声学特征的直接映射。但是模型的韵律建模功能被隐式的包含在声学模型框架中，待合成文本的韵律

信息并没有被显式建模，简单的字符表示不能表征复杂的韵律变化，因此限制了蒙古语语音合成自然度的提升。

鉴于深度学习技术显著优于传统方法的建模能力，研究人员进一步使用深度学习技术对现有蒙古语语音合成的前端模块和后端模块进行全面改进。围绕基于深度学习的蒙古语语音合成方法展开，以进一步全面提升蒙古语语音合成的整体表现。结合蒙古语语言特点和深度学习相关知识，提出了一系列创新性的解决方案，具体研究方案分为以下四个方面：

（1）融合蒙古文形态学与音系学知识的蒙古文韵律建模方法[53][54][55]：为了从模型输入和模型结构两个角度对蒙古文韵律建模的建模能力进行增强，使得蒙古语深层次上下文语义特征被充分建模，从而达到提升蒙古语语音合成的合成效果的目标。使用 LSTM 网络和 BiLSTM 网络进行蒙古文韵律建模，并充分考虑蒙古文词干后缀、音节、音素等对蒙古语单词韵律发音的影响，提出了基于词素单元的蒙古文韵律建模方法和融合形态向量和音系向量的蒙古文韵律建模方法，以提高蒙古文韵律建模在集内词和集外词的整体精度。

（2）基于多任务学习的蒙古文韵律建模方法[53]：为了进一步提升韵律建模的精度，结合蒙古文字母转音素任务与蒙古文韵律建模任务的天然高度相关性。将蒙古文韵律建模任务和蒙古文字母转音素任务整合到同一个框架，采用“编码器-解码器”网络构建多任务学习系统，使得两个任务以联合训练的方式互相学习，以期提升蒙古文韵律建模的表现。近几年，多任务学习技术已经被成功应用于诸多领域中，如机器翻译、图像标注等，该方法可以提升蒙古语语音合成模型的合成自然度。

（3）基于知识蒸馏的端到端声学建模方法[54][59]：为了克服基于端到端声学模型的蒙古语语音合成系统的天然缺陷，考虑到端到端声学模型使用自回归属性的解码器预测语音声学参数，但是由于自回归

解码器天然具有的曝光偏差问题而导致合成语音中频繁出现跳词、漏词、重复和韵律不稳定等现象。提出了一种基于知识蒸馏技术的端到端声学建模方法，提升了蒙古语语音合成系统的鲁棒性和整体表现。

（4）融合显式韵律信息的端到端声学建模方法[57][58]：为了将显式韵律信息充分融入端到端声学模型的训练过程，提出了特征级别和模型级别两种韵律信息融入方法，通过使用文本的韵律信息对端到端声学模型的训练过程进行指导，以提升基于蒙古语语音合成的整体自然度。

虽然蒙古语语音合成研究已经取得了阶段性成果，然而蒙古语语音合成技术方兴未艾，随着深度学习技术的不断发展，蒙古语语音合成研究仍然具有广阔的发展空间。我们总结了以下几点后续的研究工作：

（1）实时蒙古语语音合成系统：语音合成系统的主要任务就是将文本序列转化为接近真人发音的语音数据，而针对以语音合成服务为基础服务的上游语音交互系统来说，语音合成速度的快慢与否直接影响到上游语音交互系统的效率表现。如果合成速度太慢，就会出现语音播放卡顿或者持续等待播放等问题。因此，语音合成系统的实时性能直接关系到系统应用的效果。基于深度神经网络声学模型的蒙古语语音合成系统和基于端到端声学模型的蒙古语语音合成系统都采用逐帧解码的方式进行语音生成，这样的解码方式大大增加了合成语音的生成时间，降低了合成效率，尤其当待合成文本长度太长时间时间效率会更加低下，难以满足实时要求。因此，下一步工作中，可以以摆脱自回归解码方式的限制为研究目标，将非自回归语音合成技术应用到蒙古语语音合成模型来显著提高其合成效率、满足实时性要求。

（2）情感蒙古语语音合成系统：基于深度学习的蒙古语语音合成方法主要关注在保证发音内容传达准确的前提下如何提升合成语音的整体自然度，而没有关注语音信号中体现的情感表现力。情感语

音合成是近几年语音合成的研究热点，语音的韵律和声学特征是指导情感语音合成的主要因素。因此，下一步工作中，可以以实现情感蒙古语语音合成系统为目标。首先构建情感蒙古语语音合成语音语料库；其次，充分挖掘蒙古语语言特点，研究确定蒙古语语音的情感声学特征参数、确定蒙古语情感声学特征与情感状态的映射关系、确定蒙古文文本情感分析与场景因素结合的蒙古文语音情感预测机制等问题，实现情感蒙古语语音合成系统。

（3）多模态蒙古语语音合成系统：语音合成广泛应用于智能家居、虚拟主播、语音导航、信息播报、阅读教育、泛娱乐等领域，是人机交互的重要组成部分。人机交互的方式走过了键盘交互、触摸交互、语音交互等，每一次变化的背后都是对人和机器之间交互的便利性、自然性以及准确性所提出的更高的要求。近几年，多模态交互作为一个非常重要的人机交互方向具有势不可挡的发展趋势。首先，多模态交互能够让人类在不同的场景下选择不同的模态组合进行交互，进而提升人机交互的整体自然度；其次，多模态技术下，多个模态可以互相补充，能够通过多个模态信息的融合获得更准确的用户、情感和场景估计；最后，多模态技术能够让人机交互过程中拥有视觉、听觉和触觉等多维感觉，全方位体会机器表达的情感和语义信息。因此，下一步工作中，可以以实现多模态蒙古语语音合成系统为目标。首先构建多模态蒙古语语音合成图像语音语料库；其次，充分挖掘蒙古语语言特点，研究确定多模态蒙古语语音合成系统的多模态输入、多模态输出和中间认知环节的多模态推理和决策等问题，实现多模态蒙古语语音合成系统。

2.2.5 蒙汉机器翻译技术

围绕蒙汉神经机器翻译，从蒙汉双语语料库建设，蒙古文词切分方法，未登录词处理方法，命名实体识别方法和单语数据应用方法等

方面展开了系统的研究，取得了一定的成果。采用机器翻译和人工校对相结合的方法，构建了蒙汉翻译双语平行语料库、地名和机构名蒙汉双语词典。搭建了基于注意力的蒙汉神经机器翻译系统和基于Transformer 的蒙汉机器翻译系统。针对蒙汉神经机器翻译中的有限词典问题和蒙古文的数据稀疏问题，对蒙古文进行了切分。在蒙古文词切分方面，提出了基于 BiLSTM-CNN-CRF 模型的神经网络蒙古文词切分方法。研究了部分切分、BPE 子词切分和神经网络切分方法等不同的蒙古文词切分方法对基于 Transformer 蒙汉机器翻译的影响。研究表明，经过对神经网络词切分后的蒙古文语料，过滤掉蒙古文连接元音字母和不稳定“N”后，基于神经网络的蒙古文词切分方法在蒙汉神经机器翻译的性能比 BPE 切分和部分切分的性能好。针对蒙汉神经机器翻译的未登录词问题，我们采取基于语义相似度的未登录词替换、基于语言模型的未登录词替换和基于蒙汉对齐词典的未登录词替换方法等三种方法进行了研究。

为了缓解数据稀疏对蒙汉神经机器模型的影响，武子玉[60]提出了一种基于半监督学习的数据增强方法。该方法利用单语语料构建了无监督蒙汉神经网络机器翻译模型，仅使用蒙古语和汉语的单语语料进行训练，并利用自学习方法对两种语言进行跨语言词嵌入获得双语词典，使用双语词典和目标语的语言模型来初始化翻译模型，对语料进行迭代回译，进一步减少了模型对平行语料的依赖，缓解了机器翻译任务中的平行语料稀缺问题。同样的方法对于基于短语的半监督蒙汉统计机器翻译模型也适用。

为了让模型学习额外的蒙古语语义知识，苏依拉等[61]采用了一种多教师指导方法，进一步提高了翻译模型的输出质量。而吉亚图[62]为了获得更多的语义信息，提出了一种基于多粒度融合的单词切分方法。实验表明，字粒度对介词和独立单词具有更好的模型预测能力，词粒度更适合专有名词的预测，BPE 等子词粒度对短语具有很好的预

测效果，而词干词缀粒度则对主语等更有效。因此，为了有效的融合各种粒度的特点，他们提出了一种多粒度训练策略，即粒度过滤器（Value Iteration Network, VIN）。在进行序列编码的过程中，将句子经过不同的切分方法获取不同的粒度的序列表示，再根据 VIN 获取当前时刻对编码最有益处的粒度表示，最终获取多粒度候选序列。模型的整体结构包括三个部分：自由粒度的预处理、强化训练、价值迭代。Nier Wu 等[63]提出了一种使用 XLNet 预训练模型辅助的神经机器翻译方法，该方法缓解了 BERT 模型使用自编码方式导致的训练微调不一致问题和条件独立性假设问题，通过引入排列语言模型来以自回归的方法顺序编码序列，在获取上下文语义的同时也避免了由于使用掩模 Mask 带来的问题。

为缓解蒙汉这类低资源任务较为显著的曝光偏差问题，白天罡[64]提出了一种基于强化学习的蒙汉神经机器翻译模型。他们在蒙汉机器翻译训练过程中引入了单词级奖励和序列级奖励来平衡训练-评估过程，使得模型能够动态的观察当前迭代周期内各个单词和句子的分配信用，在下次采样时能够着重训练具有高优先级的句子经验，进而提升模型的性能。

2.2.6 蒙古文知识库

蒙古文信息处理相较于其他语言起步较晚，通过多年研究及众多专家学者的努力，蒙古文信息处理取得了优异的成果，蒙古文知识库相关研究也不例外。Google 在 2012 年提出了知识图谱（Knowledge Graph）的概念。知识图谱使用三元组的形式存储结构化数据。知识图谱的出现解决了大数据时代下结构化数据存储的难题，但随之而来的难题是用户如何获取知识图谱中存储的数据。基于知识图谱的问答系统 Question Answering over Knowledge Base，KBQA）是一种以知识图谱作为答案来源的问答系统。该系统允许用户通过自然语言获取

知识图谱中存储的数据。

近年来，随着信息技术的发展，互联网与旅游业、医疗行业的深度融合发展已成为不可阻挡的时代潮流。以往，用户想要了解旅游景点或者蒙医药相关资源时主要使用传统搜索引擎获取，但传统搜索引擎获取用户所需要的信息不够直观、冗余信息多等不足导致传统搜索引擎无法快速且准确地回答游客的问题。与传统搜索引擎不同，面向旅游和蒙医药等领域的智能问答系统会针对用户输入的问题进行语义分析且直接返回答案，这些垂直领域智能问答系统不仅可以帮助用户快速且准确地获取知识，而且还可以促进我国互联网与旅游业和医疗行业的深度融合发展，有助于充分挖掘地方特色旅游服务。蒙药种类繁多、来源复杂，是蒙古高原人民在漫长的疾病斗争历史中的思想结晶，具有悠久的历史价值与学术价值，值得我们对其进行系统性的研究。为了提供直观正确的检索效果，领域知识图谱能够为检索系统提供丰富有效的数据保障。因此，在蒙古文智能信息处理方面构建垂直领域知识图谱受到专家学者的高度关注。

1. 语义知识库

2008 年内蒙古大学研发《蒙古语语义信息词典》。该知识库中收录了 57495 个词汇，并将每个词汇的详细信息匹配到相对应的语义关系。该研究成果面向网络用户提供了 Web 版应用平台。2016 年到 2018 年期间内蒙古大学对蒙古文命名实体识别展开研究，分别使用条件随机场模型以及深度神经网络模型完成自动识别，并制定了蒙古文命名实体标注规范，建立了标注平台，形成了蒙古文命名实体标注语料库，为构建蒙古文相关知识图谱和问答系统打下基础。

在语义网络方面，内蒙古师范大学计算机学院探索本体库构建的经验以及方法，以映射的方式融合中英 WordNet 中词汇，对于无法自动转换的节点以及概念信息以手动添加蒙古文翻译的方法来构建蒙古文语义知识库，为语义相似度计算、自然语言理解等智能信息处理

问题提供基础数据。

2. 内蒙古双语旅游知识图谱

内蒙古自然资源的优势，不仅具有绚丽的山河美景，还有大草原和沙漠，是适合旅行的绝佳选择。通过构建内蒙古旅游知识图谱，使得游客对内蒙古旅游有更为直观的认识，也让想来内蒙古旅游的游客对当地景点以及这些旅游景点的信息有基本认识。但是内蒙古相关的信息太过稀缺，因此可以利用旅游网址原始数据中包含的景点及属性信息构建景点知识图谱，展示景点与属性的关系。将内蒙古各景点的旅行时间、旅行方式、景点类型、景点所属省份进行整理生成新的数据文件，将景点与景点的所有属性进行连接形成<实体 1，关系，实体 2>和<实体，属性，属性值>的三元组知识，再对三元组中的实体和属性进行蒙古文的翻译。利用本体构建工具 Protégé 来进行本体构建。定义类、属性及关系属性，连接属性关系将提前准备好的蒙古文利用 OWL 语言的等同公理完成三元组的关系链接，从而实现双语本体的构建，将本体实例化。再将数据存储到图结构的数据模型中，并在本体库中利用其扩展功能构建知识图谱。用户在选择旅游景点的时候，实际上是在筛选具有特定属性的景点。将知识图谱引入推荐系统，能够从知识图谱中的景点信息里挖掘出用户对景点属性关注的潜在信息，然后基于内蒙古景点的属性建立用户间的相似性，获得更加个性化的景点推荐，即为每个目标用户定制其可能感兴趣的景点。这样可以提高信息检索的效率。再通过用户间的相似度进行推荐，让用户使用该系统时更为个性化。基于构建好的双语旅游知识图谱，可以为游客提供旅游领域问答系统。利用蒙汉双语来构建本体，让使用该系统的人群更为广泛，对蒙古族同胞更为友好，对蒙汉双语旅游知识图谱构建、使用、推广具有较好的实际应用意义。

3. 蒙古文化知识图谱

构建蒙古文化知识图谱涉及到信息抽取，数据融合，知识表示以

及知识库构建等诸多方面。蒙古文化在千百年来世代相传，是蒙古族先民留给子孙后代的无价瑰宝，是人类文明的重要精神财富。以蒙古习俗和蒙古文化相关的书籍为来源，使用 CRFs 工具包完成命名实体识别，并且采用卷积神经网络来完成关系抽取。根据抽取的实体及属性关系，形成构建蒙古文化知识图谱的三元组。把其中抽取到的属性值进行融合，选择最优的属性值。通过查询专业书籍和利用专业蒙古文处理平台对三元组进行处理。最后，依据七步法的构建过程，利用 Protégé 工具完成蒙古文化领域本体的构建，进而实现面向蒙古文化的蒙汉双语知识图谱的构建。

1. 蒙古族历史人物信息知识图谱

蒙古族历史中涌现了大量的杰出人物和事迹，史料文献丰富，构成了一个庞大的知识体系。通过收集挖掘蒙古族历史人物信息，构建人物及其相关事件的知识图谱，可以为历史爱好者和研究者检索和研究提供帮助。蒙古族历史人物相关知识图谱尽可能地列出涉及的重要概念以及术语，定义了人物、事件、部落、书籍等四个实体类，并定义了实体相关属性以及实体间的语义关系。蒙古族历史人物相关语料经过蒙古文编码归一化等处理后进一步进行实体对齐、属性值融合等处理，并导入到 Neo4j 图数据库，从而完成蒙古族历史人物信息知识图谱的构建。

2. 《蒙古秘史》全文检索平台

蒙古语词汇语义网在语料库语言学或文本挖掘方面的一个典型的应用是《蒙古秘史》全文检索平台（www.mnuuts.com）。该平台由呼和浩特民族学院计算机科学与信息工程学院金罡副教授率领的团队研发完成。该平台依托国家社科基金“蒙古语 TEI 标注模式历时语料库的建设与应用研究”（12XYY026）、教育部人文社科基金“《蒙古秘史》词汇语义树库构建研究”（19YJA740023）等项目，将《蒙古秘史》中名词、形容、动词以同位关系、上下位关系、整体与部分

关系组织为树形结构数据库，并将此数据库融入到《蒙古秘史》全文检索平台，为《蒙古秘史》语料库检索增加了语义检索功能。

传统的语料库检索主要以字词（字符串）或固定模式（正则表达式）来检索语料，没有涉及到语义的检索。《蒙古秘史》全文检索平台中的语义检索功能从词与词之间的语义关系全面搜索语料内容，从而达到了挖掘《蒙古秘史》中的知识的目标。例如，《蒙古秘史》中有多少地名出现、有多少植物名出现等问题轻而易举地回答。图 2-9 为《蒙古秘史》全文检索平台中利用语义检索功能查找所有植物名词的结果。结果中的每一项内容也可用树形结构图来查看该项的语义层次关系。如图中的 saqal bayyan(撒中合勒伯顔)一词为例，从可视化结构图可容易地得知该单词是植物类中的草类，也可在树形图中查看与该词有同位关系的、上下位关系的其他词（如图 2-10）。

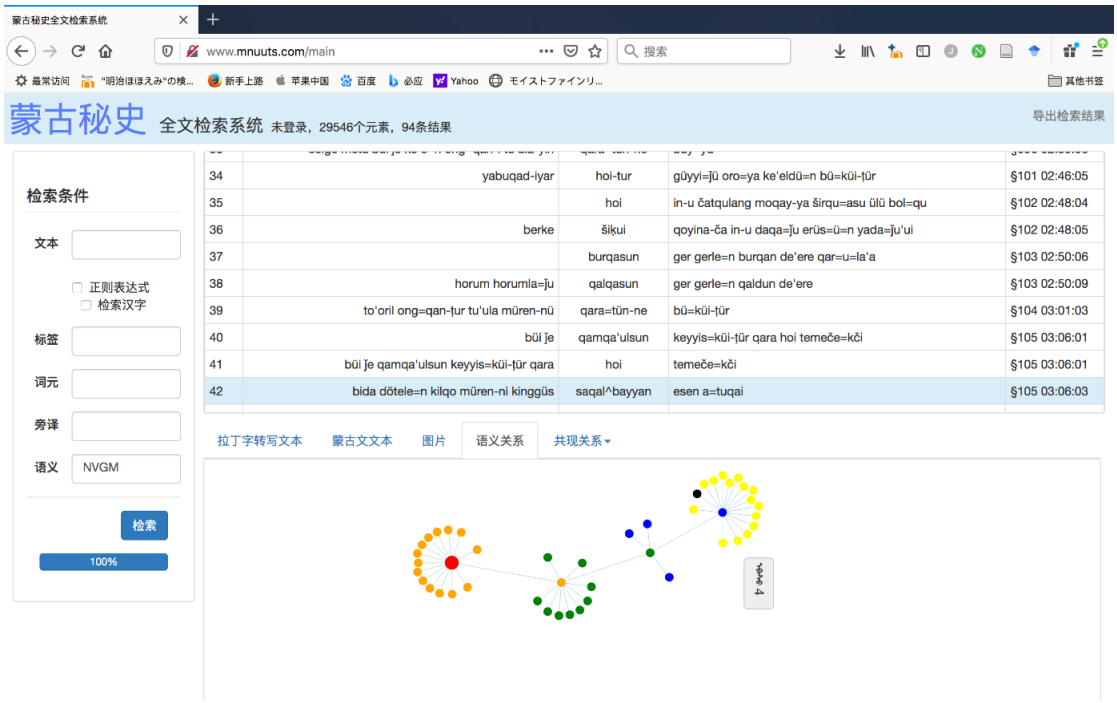


图 2-9 语义检索

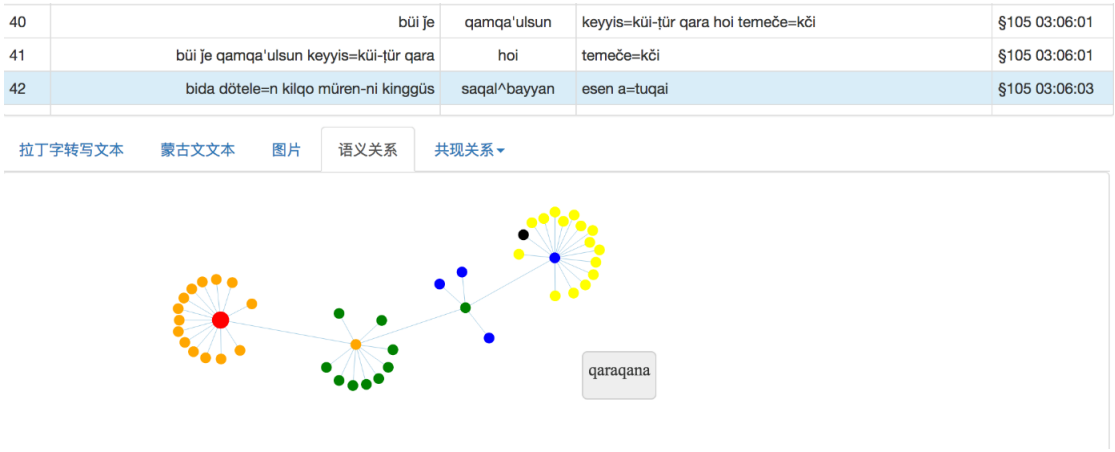


图 2-10 语义关系

3. 蒙古语五畜知识平台

蒙古语词汇语义网在语料库语言学或文本挖掘方面的另一个典型的应用是蒙古语五畜知识平台（www.wuchu.com）。该平台由呼和浩特民族学院经济管理学院布音其其格副教授的团队研发完成。该系统的研发依托国家自然科学基金资助项目“蒙古语词汇语义网研究”（项目批准号：61363053）、内蒙古自治区高等学校科学技术研究项目“阿拉善骆驼词汇语义数据库构建”（项目批准号：NJSY226）、内蒙古自治区民委蒙古语言文字科研资助一般项目“蒙古族五畜词汇语义数据库构建”（项目批准号：MW-YB-2015034）、内蒙古自治区民委蒙古语言文字科研资助重点项目“蒙古族五畜相关民间文学知识库建设”（项目批准号：MY-ZDI-2018001）等项目，实现了五畜相关词汇及民间文学作品的搜集整理入库及查询维护管理功能。

蒙古语的 ᠲᠠᠪᠤᠨ ᠬᠤᠰᠤᠨ ᠮᠠᠯ（“五畜”）是指蒙古高原上主要经营的 ᠮᠣᠷᠢ（马）、ᠲᠡᠮᠡᠭᠡ（骆驼）、ᠦᠬᠡᠷ（牛）、ᠬᠤᠶᠢ（绵羊）和 ᠶᠠᠮᠠᠨ（山羊）。蒙古语中五畜相关词汇非常丰富，比如关于五畜的岁数、性别名称、用具的名称、身体部位名称以及畜产品名称等相关五畜文化的词汇规模很大，并且区分程度也是非常细致。本系统搜集了五畜相关词汇，然后对词汇进行语义分类并完成词汇语义解释，常用词汇使用下位语

义场方式提供了说明，最终以数据库形式存储了语义信息，旨在使五畜相关词汇较全面、系统、正确的保存，为使用者提供快捷方便的词汇语义查询学习服务。系统中除了五畜相关词汇以外还增加了五畜相关民间文学作品，例如谜语、谚语、神话故事、赞颂词、民歌歌词等内容，使系统中的知识更加丰富，后续将不断扩充完善数据，通过动态更新满足各类用户的知识学习使用需求。

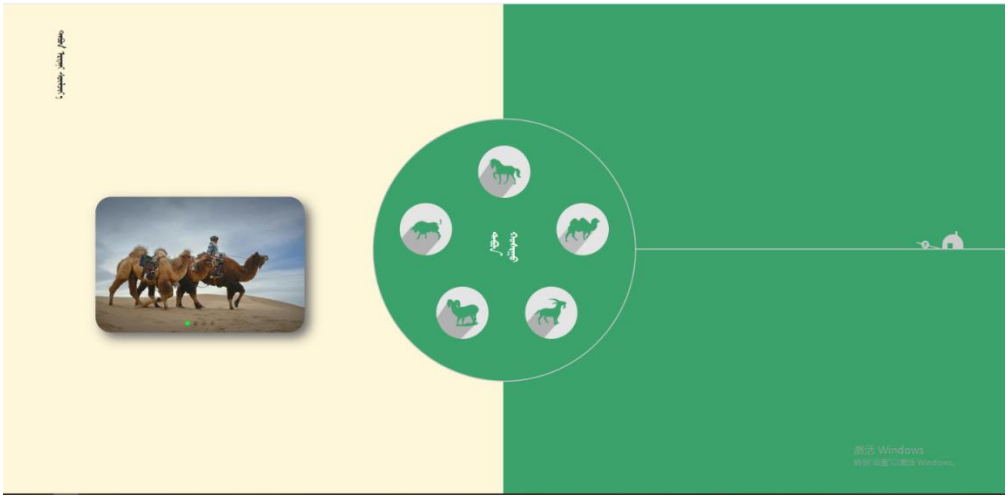


图 2-11 五畜知识网站首页

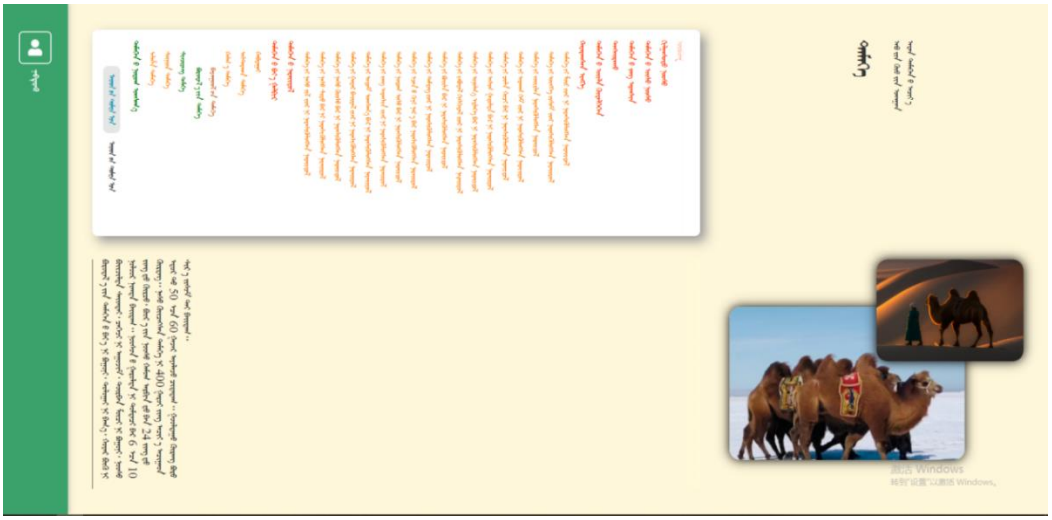


图 2-12 骆驼相关词汇语义层级关系

2.2.7 蒙古文语义挖掘技术

2.2.7.1 蒙古文命名实体识别相关研究

命名实体是文本中基本的信息元素，是正确理解文本的基础。命名实体识别 (Named Entity Reeognition，NER)的主要任务是识别出文本中出现的名字实体和有意义的数量短语并加以归类，主要包括人名、地名、组织机构名、时间表达 式、日期、数字表达式等。命名实体识别是自然语言处理中的热点问题和基础性 工作，对自然语言处理具有极其重要的意义，并被应用到自然语言处理的许多领域，如信息检索、信息抽取和机器翻译等。

蒙古文命名实体识别研究起步比较晚，大多采用统计机器学习或神经网络的方法识别蒙古文人名、地名和机构名。人名和地名的识别中 CRF 方法[65-70]被广泛采用，封闭测试准确率达 90%以上。随着神经网络技术的发展，采用深度神经网络框架提出基于蒙古文词向量的命名实体识别方法，进行蒙古文命名实体识别研究。基于蒙古文词向量的方法[71]、蒙古文词素向量的方法[72][73]、融合语言模型和注意力机制的方法[74]、基于 transformer 神经网络模型[75]、基于 Bert 词嵌入模型[76]等方法被提出，并获得较好的识别性能。

2.2.7.2 基于蒙古文 Bert 词嵌入的蒙古文实体抽取

词向量是自然语言处理任务的基础。跨语言词向量借助迁移学习将单语词向量映射到一个共享的低维空间，在不同语言间进行语法、语义和结构特征的迁移，可以对跨语言语义信息进行建模，是解决低资源语言信息处理和语言鸿沟引起的跨语言信息处理的重要基础环节。然而目前跨语言词向量的学习性能较大程度上依赖于大规模的平行语料或高质量的种子词典等，对于平行语料较少的蒙汉跨语言词向量的学习效果不太明显。蒙古语为低资源语言，很难获取大规模蒙汉平行句对，而且蒙古语构词方法独特而复杂、形态多变的粘着性，使用神经网络学习蒙汉文跨语言词向量时导致了更严重的数据稀疏和

集外词问题。多语言 BERT 模型是在大规模高资源语言语料上预训练的一种动态词向量模型，其中包含了丰富的多语种语法、语义信息，在此基础上二次精调模型能够解决集外词和数据稀疏问题。但它没有考虑蒙古语的构词法，预训练时也没有训练蒙古语文本。针对以上问题，主要研究内容包括以下几点：

（1）针对目前只有蒙古文静态词向量学习模型，提出一种深度迁移蒙古文动态词向量学习模型。在多语言 BERT 预训练模型的基础上利用小规模的语料进行精调，通过迁移学习将高资源语料上学习的语法、语义特征映射到低资源的蒙古语动态词向量表示。为验证方法的有效性，在团队构建的数据集上用不同的模型进行了同义词对比实验，并利用 K-means 聚类算法对蒙古文词语进行聚类分析，最后在嵌入式主题词挖掘任务中进行了验证。实验结果表明，BERT 学出的词向量质量高于 Word2Vec 静态模型，相近词的向量在向量空间中的距离非常近，不相近词的向量较远，在主题词挖掘任务中获取的主题词有密切的关联。

（2）针对构建蒙汉跨语言词向量缺乏大量平行句的问题，提出一种深度语言知识共享迁移模型。利用小规模的蒙汉平行句对，通过共享模型参数、语言知识联合学习跨语言词向量表示、子词（字）学习以及平行句的判断。并通过自注意力机制进一步学习蒙汉句对中每个词之间的语义关系，从而构建蒙汉跨语言词向量。最后，通过双语词典归纳任务评估训练的蒙汉跨语言词向量对齐性能，表明提出的方法与基线模型相比有明显的提升。

2.2.7.3 融合数据增强和知识迁移策略的蒙汉跨语言知识抽取

知识抽取是以构建知识图谱为目标，从非结构化文本中抽取“实体-关系-实体”结构化三元组知识单元的任务。主要包括命名实体识别、实体间关系抽取等信息抽取关键子任务。伴随着全球化的深入，不同语言、不同领域的知识共享与联系日益紧密，从多语言文本大数

据中进行跨语言知识抽取的需求越来越广泛，成为文本挖掘技术研究热点。蒙古语使用者主要分布于中国内蒙古地区、蒙古国和俄罗斯。近年来，随着蒙古文数字化资源日益增多，对蒙、汉语文本进行跨语言知识抽取进行分析挖掘、构建领域知识图谱的需求越来越广泛：例如构建蒙汉语旅游知识图谱、中医（蒙医）医药学知识图谱、发现蒙汉语新闻热点主题、分析比对蒙汉语论文及著作的研究主题关系和演变趋势、检测蒙汉语文本的相似性、建立蒙汉语医疗健康智能问答系统、建立蒙汉机器翻译、跨语言检索与推荐系统等。该研究对于解决民族地区蒙汉语言差异引起的语言鸿沟，实现国家通用语言与地方民族语言的知识共享和利用、传承优秀民族传统文化，促进民族地区乃至“一带一路”地区经济文化繁荣和发展、铸牢中华民族共同体意识都具有十分重大的现实意义和广泛的应用前景。

近年来，基于深度学习的各种信息抽取模型，被广泛应用于自然语言处理领域，并取得了很好的效果。针对跨语言实体识别、实体间关系抽取的模型，研究者普遍采用的方法是：首先通过人工或机器翻译构建大规模双语平行语料或双语词典，然后通过双语句对齐或词对齐的显式监督信号学习其中隐含的跨语言映射关系，得到共享语义空间表示，并在共享语义空间中利用带实体、关系标注的数据进行抽取模型的有监督训练。该方法应用在具有大量标注数据的英语和一些欧洲语言时，取得了很好的效果。但在数据稀缺的小语种语言中，这种方法往往难以取得较好的结果，依然面临以下三个问题：

（1）蒙古语属于低资源语言，仅利用已有的小规模蒙汉平行语料和带标注蒙古文语料监督训练跨语言实体和关系抽取模型无法取得好的性能。

（2）蒙古语构词形态变化复杂，因形态变化派生的新词会导致蒙古语词表的规模变得非常大。把蒙古语词汇整体作为处理单元会损失掉其中包含的大量语法、语义信息。

（3）基于神经网络的实体和关系联合抽取，传统上采用有缺陷的流水线模型。流水线模型没有建立实体识别和关系抽取两个任务之间的依赖，存在错误传播、信息冗余等问题，最终影响“实体-关系-实体”三元组知识单元的抽取效果。

针对以上问题，本文从小数据深度学习和无监督深度学习方法入手，研究一种利用蒙汉非平行语料和蒙古语无标注数据的跨语言映射数据增强和跨语言模型迁移增强无监督深度学习方法，把高资源汉语的标注数据和训练好的模型知识都高质量地迁移到低资源蒙古语模型上，降低深度学习对低资源语言数据的依赖，并构建比现有更优性能和跨语言适用性的相关模型和算法，达到更低的训练成本。在此基础上进一步研发的标注工具软件，既可以用其建立新闻、医药相关领域的标注语料库，又可以用于构建蒙汉领域知识图谱的知识抽取任务中。具体工作内容如下：

（1）蒙汉子词（字）和句向量联合跨语言映射数据增强的蒙古语实体识别：研究一种基于预训练语言模型的蒙汉子词（字）向量和句子向量的跨语言语义空间映射的多策略对抗学习方法。

（2）基于跨语言知识迁移学习的蒙古语命名实体识别、蒙汉跨语言实体识别：借鉴基于教师-学生网络的知识蒸馏思想，将训练好的汉语模型作为教师，其输出作为软监督信号，指导蒙古语学生模型的训练。

（3）基于参数共享的蒙汉跨语言实体和关系联合抽取：利用跨语言预训练模型将实体识别和关系抽取两模型的编码器完全融合进行参数共享，在顶层解码阶段采用各自分类器获取实体和关系信息。

（4）研制蒙汉跨语言实体和关系联合标注工具软件。

2.2.7.4 多语课程知识图谱及其跨语言检索

结合地方语言特色和师范院校的教师教育特色，开展多语言知识图谱构建研究，构建了多语课程知识图谱。

知识图谱是人工智能技术突破性进展的重要组成部分。作为“人工智能+教育”驱动下的教育新业态，智能教育必然与知识图谱存在着紧密联系。借助教育知识图谱，重构知识之间的链接关系，形成学科专业的网状知识体系结构，能够有效支持灵活、精准“教”与个性、终身“学”。

依据智慧教育与课程知识图谱之间的映射关系，面向智慧教育的课程知识图谱构建设计采用自顶向下的方式构建课程知识图谱，构建框架如图 2-13 所示。

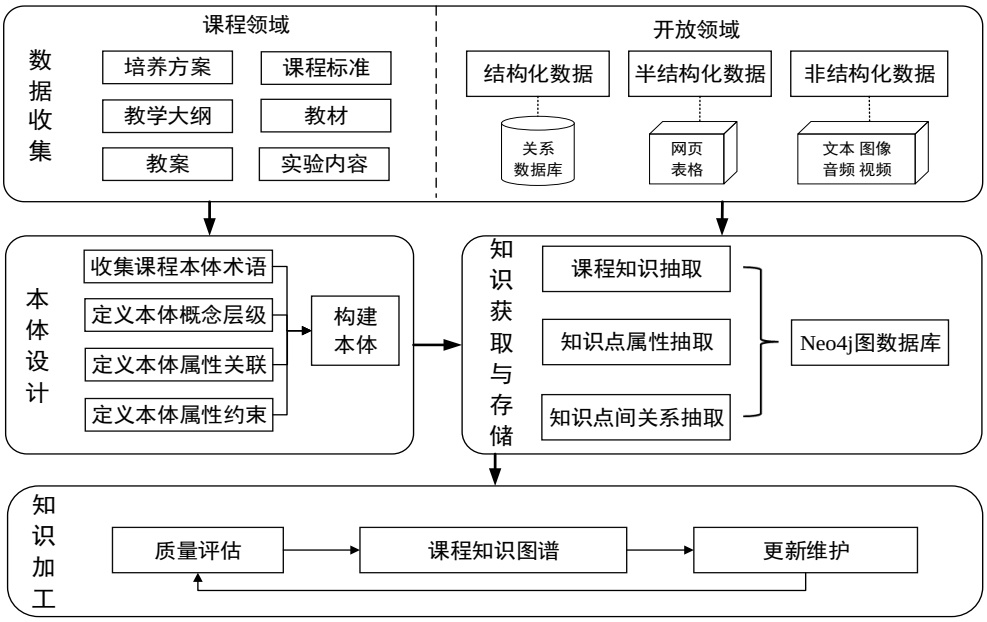


图 2-13 面向智慧教育的课程知识图谱构建框架

在该框架中，课程知识图谱构建过程分为数据收集、本体设计、知识获取与存储、知识加工四个阶段。具体地，丰富的数据为本体设计和知识获取与存储两个阶段提供坚实支撑；本体设计为知识获取与存储阶段提供了理论基础和抽取原则；知识获取与存储阶段在本体的指导下从数据中抽取相关课程知识与关系，初步形成课程知识图谱；知识加工阶段通过质量评估与更新维护得到高质量知识图谱，为智慧环境教育的创造提供条件。

2.3 藏文智能信息处理

在实现中华民族伟大复兴的第一个百年，藏文信息处理在以通用语为核心的中华民族统一多平台支持下，研制了一系列具有自主知识产权的民族文字软件产品。践行了五个认同，铸牢中华民族共同体意识，使我国的少数民族同伟大的祖国一起进入信息化社会，完成了一个历史阶段的重大使命。

2.3.1 藏文信息处理基础理论研究

2.3.1.1 藏文词法分析

藏语词法分析工作主要包括：1）藏语紧缩词识别，代表工作为西北民族大学提出的基于序列标注方法的紧缩词识别方法，是目前主流的方法，以及青海师范大学提出的还原算法。2）藏语分词模型，以序列标注模型为主，如西北民族大学提出的统一的分词、紧缩词标注模型，以及青海师范大学提出的基于词性约束的藏文分词方法。3）藏语命名实体识别，典型代表为西北民族大学提出的统计与规则相结合的方法，以及西藏大学提出的融合音节特征的藏文命名实体识别方法。4）分词评测，2017 年，在中文信息学会举办的第一届藏语分词评测中，西北民族大学获得第一名，准确率达到 92.9%，2021 年第二届藏语分词评测第一名由青海师范大学团队获得。5）开源藏语分词系统，藏语分词研究相对较为成熟，西北民族大学开发的开源分词系统得到广泛应用，地址：https://github.com/liyc7711/tip-las。

2.3.1.2 藏文句法分析

藏文句法分析围绕依存句法分析和短语句法分析两方面展开。

国内高校、科研院所主要以藏文依存句法分析研究为主，代表性工作包括西藏大学的祈使句依存分析；西北民族大学基于依存关系的藏语句法分析、藏语状态词研究，藏语词法句法联和分析研究；青海师范大学基于词对依存分类树库构建、藏语依存树库构建、基于深度

学习的藏语依存句法分析等工作。目前文献中藏文复合句依存句法分析的准确率达到 88.72%，藏文疑问句句法分析准确率、召回率和 F 值分别达到 96.0%、96.1%和 96.1%。

藏文短语句法分析研究处于起步阶段。

西北民族大学构建了藏语短语句法树库，应用于藏文信息抽取以及藏汉机器翻译研究工作中。进行藏文短语句法分析研究，融合更多层次的语言学线索，体现更多的语言学特征。为了方便研究藏文信息处理的科技工作者能够更好地理解藏文，创新性地提出句法树的叶子结点本身为藏语词，通过词对齐技术可以为藏语词接入对应的汉语翻译，形成新的句法树。

2.3.1.3 藏语语义分析

国内相关高校和科研机构在藏语语义分析领域的工作集中于词汇级和句子级语义分析领域，篇章级语义分析成果较少。藏文语义分析围绕藏文语义依存展开，代表性工作包括西北民族大学制定了基于依存关系的藏文语义角色标注体系；西藏大学结合藏文语法体系，设立符合藏文语法体系的依存标注关系体系，并设计了基于判别式的句法分析算法；青海师范大学采用基于判别式模型，提出不同句型的藏文依存句法分析方法。

西北民族大学在汉语浅层语义分析领域也有相关成果，基于传统机器学习方法，提出基于条件随机场的语义标注模型改进方法，在训练过程中融入词法及句式等多层级语言学线索，有效增强了标注性能；改进了深度学习模型架构，优化序列标注模型性能，还探索了对多特征的采样提取，进一步扩展模型潜力；同时对序列标注模型进行结构优化升级，应用于汉语语义角色标注任务。

2.3.2 藏文信息处理应用研究

2.3.2.1 藏汉机器翻译

藏汉翻译工作主要包括：1）藏语词语/亚词切分方法，典型代表

为西北民族大学提出的基于音节的藏语亚词切分方法，以及北京理工大学提出的多策略切分粒度处理方法，无监督的亚词切分方法（SentencePiece）也得到成功应用。2）针对藏汉翻译语料资源不足问题，西北民族大学提出了基于模型迁移的藏汉神经机器翻译方法，青海师范大学提出了基于迭代式回译策略的藏汉神经机器翻译方法和融合单语语言模型的藏汉机器翻译方法。3）藏汉翻译语料库建设，藏汉翻译平行语料相对较少，2022 年以前，全国机器翻译大会（CWMT）的机器翻译评测的训练语料规模在 20 万句左右。在国家重点研发计划支持下，西北民族大学为 CWMT 2022 藏汉翻译评测提供了 100 万句对的训练集，极大缓解了藏汉翻译数据规模不足问题。4）在最近的 CWMT 2022 机器翻译评测中，所有参赛系统均采用神经机器翻译模型，其中成绩最好系统的 BLEU 达到 33.95，表明藏汉机器翻译质量整体上达到了较高水平。

藏汉机器翻译中，所采用的模型与国内外前沿研究保持基本同步，东北大学、西北民族大学、青海师范大学、西藏大学等单位已经推出藏汉机器翻译相关产品可供应用。

2.3.2.2 藏文文本分类

当前国内开展藏文本分类相关研究的单位主要有：西北民族大学、复旦大学、青海师范大学、中央民族大学、西藏大学等高校。

西北民族大学研究团队自 2010 年左右开始研究藏文文本分类的相关研究。研究团队已构建出藏文新闻文本分类和藏文评论文本情感分类实验用语料库，藏文文本分类主要通过 TF、TF-IDF 等方法进行特征提取，然后采用主成分分析法等方法进行特征降维，最后通过逻辑回归、朴素贝叶斯、支持向量机、K 近邻等传统机器学习方法，开展文本分类研究。随着深度学习在自然语言处理领域的广泛应用，将 CNN、LSTM 等深度神经网络融入藏文文本分类中开展相关应用，并取得的不错的分类结果。

为了解决以词为文本表征对象的分类模型中低频词条大量存在现象和藏文词切分效果并非理想等问题，提出了以藏文音节为文本基本表征单元的方法，基于深度学习模型强大的多层非线性处理单元级联方式进行特征提取，最终在藏文新闻文本分类和藏文评论文本情感分类的准确率方面均优于其对应的基于词条模型。

2.3.2.3 藏文知识图谱

西北民族大学研究构建的藏文知识图谱，旨在描述客观世界的概念实体、事件及其之间的关系，能够从语义层面上描述各类实体与概念以及概念之间的关系，进一步提升信息抽取精度。综合利用互联网、各种自建知识库、行业语料库等多种数据源，获取术语、概念和之间的层级关系等藏文图谱元数据知识，提高抽取藏文命名实体和实体关系的准确度与覆盖面；结合藏文语义网络完成实体属性补全，融合藏文句法属性与深度学习实现不同知识库之间的实体对齐，构建完备的藏文知识图谱。当前，西北民族大学在藏文实体、实体关系、概念抽取、实体属性补全和实体对齐等藏文知识图谱基础研究，取得了一定的进展。

2.3.2.4 少数民族语言域名国际化状况及藏文域名国际化

藏文域名国际化工作处于起步阶段。Tibetan Script GP 小组 2017 年在中国互联网信息中心成立，中国有 12 人（其中西北民族大学两人），不丹有 6 人，旨在推进藏文域名国际化工作。西北民族大学作为藏文域名国际化工作组（Tibetan Script GP）的参与单位之一，将参与研究生成域名规则集，并形成报告提交 ICANN 组织，最后将藏语文字域名加入国际化域名生成规则集。完成了国家民委委托《我国少数民族语言文字域名国际化现状调查研究》的研究工作，相关成果得到了有关部门的认可。

2.3.2.5 藏语舆情分析

西北民族大学针对公安机关、安全部门和有关政府部门无法实时

监控印刷品、存储介质和互联网藏文舆情动态监测的迫切需求，在甘肃省科技重大专项项目“面向社会安全的藏文舆情云分析平台研究”的支持下，针对藏文命名实体识别、藏文文本智能倾向性分析、藏汉辅助机器翻译、藏文舆情云分析服务平台等方面进行深入研究，开发了多载体的藏文舆情分析云系统平台，并将平台应用于公安、安全等部门。 随后与中央民族大学联合承担国家科技支撑计划子课题“少数民族网络舆情综合分析与云服务关键技术研究及应用示范”，系统化产出藏文网络舆情分析报告，为应对社交媒体和短视频舆情分析的需要，在国家自然基金的支持下不断将社会网络分析、多模态藏语处理技术应用于藏语舆情分析领域。

此外，西藏大学集成词法分析、机器翻译、跨语言检索等技术成果，研发了跨语言互联网藏文舆情分析技术及其软件系统。青海师范大学将“互联网+藏语信息处理平台建设项目”进行成果转化，融合藏语网络信息采集、藏文搜索引擎、双语机器翻译、文本情感分类等技术，实现藏语舆情分析及应用子系统。中央民族大学在国家自然基金重点项目“跨语言社会舆情分析基础理论与关键技术研究”的支持下进行了藏语社会舆情分析的关键技术研究。

2.3.2.6 藏语言资源监测

2010 年，西北民族大学与中央民族大学合作共建“国家语言资源监测与研究中心少数民族语言分中心藏语文研究基地”(以下简称：藏语文研究基地)，其主要任务是对藏语文的主要媒体，包括平面媒体，如报纸、期刊、图书、教材以及网络媒体和有声媒体的语言资源进行采集、加工、整理进入国家少数民族语言动态流通语料库，进行藏语文的动态监测与研究，并将研究成果提交国家语言文字工作部门发布，为国家民族语言政策的制定和调整提供参考。

十多年间，藏语文研究基地坚持开展文化共同体意识研究、藏语言文字和谐研究及藏文信息处理，分别于 2010 年、2011 年及 2012

年主持撰写了面向小学和中学藏语文教材、平面媒体及网络媒体中藏文词汇使用状况的调查报告，被国家相关部门采纳，收录于三个年度的《中国语言生活绿皮书 中国语言生活状况报告》。

2.3.2.7 藏文情感分析

西北民族大学开展面向藏语环境的情感语义分析，针对带有情感色彩的主观性信息计算、推理情感的倾向性。在单文本模态基础上融合视频图像、语音等多模态信息，从情感产生的内在机理和社会因素两个角度，运用多种技术对不同模态下的情感信息进行描述和融合，丰富情感度量的上下文语义，探察细微情感表达的复杂程度，从多模态角度进行情感信息的识别和理解。研究成果为藏语语境下的个人、社团群体观点监测、兴趣挖掘、行为评价等提供了技术支撑，加深对藏族网络用户的思想情感了解，掌握情感趋势，对保障藏族社区网络生活和信息化安全具有重要的科学价值和社会意义。

2.3.3 机器视觉与听觉

2.3.3.1 藏语语音识别

西北民族大学自 2006 年开始藏语语料库和发音词典构造等工作。2010 年训练了基于决策树的藏语拉萨话三音子模型，确定了一套完整藏语大词表连续语音识别的完整方案，实现了基于 HMM 的单说话人拉萨话连续语音识别，音节错误率达到 7.8%。2018 年采用 Lattice-free MMI 准则的 TDNN-HMM 声学模型和迁移学习，使大词表说话人无关的拉萨话识别的音节错误率降达到了 14.74%；同年，使用基于 LSTM-CTC 的端到端拉萨话语音识别上实现 18.71%的音节错误率。2020 年在 DNN-HMM 多任务学习声学模型上词错误率有 1%~2%的相对减少。2021 年探索了基于少部分配对语音文本和大量未配对的语音本文的半监督方法在藏语语音识别上的效果，期望解决藏语语料少的问题；同年研究藏语安多话语音识别，使用 DNN-HMM 的方法分别取得了音节错误率 14.07%，应用端到端的方法音节错误

率降低到了 5.3%。2022 年探索 wav2vec 特征在藏语语音识别上的应用。

目前国内还有天津大学、中央民族大学、西北师范大学、青海师范大学等高校及腾讯、科大讯飞等公司在研究藏语语音识别。由于资源相对稀缺，因此藏语语音识别的研究主要集中在选择声学模型建模单元、纳入有效音调信息以及基于 LFMMI、迁移学习和变分建模单元的藏语语音识别系统的改进。天津大学团队提出使用迁移学习的多语种端到端语音识别方法取得了相对字错误率 14.2%的性能提升。

2.3.3.2 藏语语音合成

西北民族大学自 2006 年左右开始研究藏语语料库的构建、藏语语音合成前端文本处理中的藏语的字素转音素、音素切分、发音词典构造与文本正则化。在 2010 年左右实现基于统计参数的藏语语音合成，使用 HMM 传统语音合成方法在藏语拉萨话上达到了 3.83MOS 打分结果。在 2019 左右开始采用深度学习中序列到序列的方法实现藏语语音合成，采用声学模型与声码器级联的方法将藏语经过简单的预处理后送入，经过模型推理就可以得到语音波形，实现 3.92MOS 打分结果。在近几年随着模型复杂度的增加，藏语拉萨话语音合成的完整度自然度逐渐提升至于人类发音接近的水平，使用 Tacotron 和 Deepvoice3 模型都达到不错的效果，后采用 Tacotron2 与 waveglow 模型实现了 4.00MOS 打分结果。

当前国内还有中央民族大学、西藏大学、西北师范大学、青海师范大学，这些高校参与藏语语音合成相关研究。由于藏语文本特殊的二维机构，难以使用字符直接参与合成训练，对于藏语语音合成普遍采用音素序列进行训练，青海师大团队提出将文本拆解为字丁的方法进行合成。

2.3.3.3 藏语语音视位

西北民族大学立足于藏语智能信息处理学科优势，引入国际先进

的高精度三维面部捕捉系统并将其应用至藏语语音视位研究中，构建了较大规模且具备科学性和可用度的高精度藏语三维视位数据库，实现了语音视位研究方法从二维到三维的创新型转变。

采用深度学习方法对大量藏语三维视位数据进行训练分析，实现了语音/文本驱动的藏语三维视位动画合成，对人脸动画、可视语音合成、机器唇读、数字娱乐等领域有着重要的推进作用。

2.3.3.4 文字识别

文字识别是利用计算机自动识别字符的技术。

我国多文种文字信息技术是在国家自然基金重点项目“多民族文字识别及理解的理论与方法研究”支持下，丁晓青教授主持，由清华大学、西北民族大学、新疆大学、内蒙古大学联合攻关。经历文字信息采集、信息分析与处理、信息分类判别几个阶段，从印刷体识别到手写识别、从文字识别到版面识别，历时六年，完成了“统一平台蒙藏维哈柯朝主要民族文字汉英混排文档综合识别理解系统”。系统实现了多民族文字对汉语的可理解翻译。多民族文字识别从实验室技术到产品转变，已经进入行业应用成熟阶段，为智能信息处理奠定了良好的技术基础。目前，在工业界应用领域的少数民族 OCR 识别软件印刷字符的单字识别率达到 98%以上，少数民族 OCR 识别软件对实际文本的识别率平均达到 95%以上。

2.3.3.5 藏语虚拟现实

在元宇宙的新型数字科技背景下，为深入挖掘藏文宝贵文化资源，深化藏文文化和新科技融合发展，西北民族大学立足科技与文化交融点，以虚拟现实、增强现实、虚拟数字人等一系列元宇宙底层技术链为依托，融合三维运动捕捉技术、虚拟现实技术、虚拟数字人生成算法等关键性技术，构建了高精度国家通用语/藏语三维视听数据库资源。

在虚拟现实与虚拟数字人研究方面，为了提高人脸发音动作数据

精确度，西北民族大学将国际领先的三维运动捕捉设备运用到国家通用语/藏语多语言虚拟数字人生成中，并结合藏文文本、音频、视频等辅助国家通用语/藏语三维虚拟数字人研究，形成了以通用语/藏语语音视位研究、多语言虚拟数字人生成、虚拟现实背景下的藏语语言传承与保护为主线的研究体系，将藏文文化单一的传承与保护逐渐转变成多元的信息化传承方式。

在此基础上，将研究深入推广到民族多元文化中，产生了藏族锅庄舞数字化、敦煌舞三维教学演示系统等一系列多民族文化传承与保护的优秀研究成果，致力于打造集多民族语言、艺术文化于一体的虚拟数字人，构建丝绸之路多语言民族舞蹈国际传播平台，提升民族特色文化的传播度与竞争力，促进多元民族文化传承与创新发展。

2.3.4 语音多模态

西北民族大学中国民族语言文字信息技术教育部重点实验室在语言和口传文化声学特征研究、语言与口传文化生理机制研究、语言与情感脑认知机制三个领域开展特色创新工作。

2.3.4.1 语言和口传文化声学特征研究

建立藏语（安多话、拉萨话）、蒙古语、维吾尔语、东乡语、保安语、裕固语、甘肃方言和各民族学习汉语普通话的的语言声学数据库，主要开展多语言语音声学模型构建、多语言语音格局分析、多语言韵律特征研究、民族学生汉语普通话学习的声学研究、汉语方言的声学特征等研究工作。建立了藏族格萨尔说唱、蒙古长调、藏族民歌、裕固族民歌、花儿民歌等 5000 小时以上声学数据库。主要通过语音特征进行分析，从科学层面解释了民族语口传文化的发音机制和言语生理机制。为传统语音学研究提供了新方法、新理论，为言语工程提供了数据支持。2017 年作为技术团队完成中国语言资源保护工程，按照国家标准建立了甘肃 27 个方言点和临夏花儿、秦州小曲、武都高山戏等 20 余种口传文化语音视频数据库。

2.3.4.2 语言与口传文化生理机制研究

在该领域开进行嗓音、动态腭位、二维唇形视频、三维唇形、气流气压、呼吸带、舌线超声等数据采集工作。重点开展汉语方言，东乡语、保安语和裕固语等稀有民族语言的生理研究，用多模态数字化方法进行传承研究，并通过建立语音生理多模态资源库方式对语言资源进行保护和发声机制研究。

2.3.4.3 语言与情感脑认知机制研究

以各语言的语音特征感知、词汇阅读加工、多语人语义理解和不同文字系统阅读为研究对象，借助脑电信号采集、眼动轨迹记录和行为测试技术，以语语言学、认知语言学、心理学和神经科学等理论为基础，探究语言加工的大脑功能分区、眼动轨迹，认知模型和认知神经机制。该工作是推动语言理论创新和人工智能发展的前沿性研究。目前已经开展了藏语多语人词汇识别、蒙古语元音和谐感知、侗语声调感知等方面的工作。

2.3.5 多语言信息处理产业化

2.3.5.1 国家通用语/藏语远程教育平台

西北民族大学研发的国家通用语/藏语远程教育平台，是国家科技支撑计划项目“少数民族语言文字信息处理共性关键技术研究与示范应用”的科研成果，国内首家大型国家通用语/藏语远程教育门户网站。

开发了用于非汉语母语者通用语学习系统，包含九年义务教育网络教学资源，大学本科基础课和专业课网络教学资源，农林牧医科普教育网络教学资源等远程教育资源库。

通用语/藏语远程教育平台，极大提高教学水平，实现地区教育的快速发展，缩小藏区与内地之间的差距，提高藏区人民的教育、文化、科技素质，缩小民族差别和从整体上地区差距，推广通用语教育，从而在根本上铸牢中华民族共同体意识。

2.3.5.2 “一带一路”特色农产品多语言电子商务平台

西北民族大学的“一带一路特色农产品多语言电子商务平台”来源于国家科技支撑计划课题“民族特色农产品多语言网络交易展示平台关键技术集成与应用示范”。

该平台是国内外第一个以汉语为核心的多语言平行对照电商平台，以“一带一路”沿线特色农产品为特定交易对象的电商平台，平台发挥西北民族大学民族语言学科优势，瞄准民族地区商贸交流存在语言障碍、信息服务薄弱、农产品销售困难的现实困境，尤其是聚焦目前国内外缺乏民族语言电子商务平台的瓶颈制约，带动农村贫困地区经济发展。平台自 2019 年起，有 500 多家企业受益，3000 余种特色农产品通过汉语/英语/藏语/蒙古语/维吾尔语五种语言在网销售。

该成果在 2019 年、2020 年多次被甘肃省省长通过批示和专题会议的形式推进项目转化落地。并将此平台建设作为重点工作写入甘肃省政府 2019 年工作报告，2019 年 11 月 27 日被省政府列入《新时代甘肃融入“一带一路”建设打造信息制高点实施方案》。

2.3.5.3 非汉语母语者国家通用语推广与普及

西北民族大学研究语言习得理论和认知语言学对国家通用语推广的指导意义；国家通用语的知识表示体系；国家通用语多层次特征的数字化模型与算法；国家通用语多模态记忆理论与方法；开发国家通用语的智能教育系统。用以解决推广普及非汉语母语者学习国家通用语语言文字本体的差异，语言知识构成差异

在中央民族工作会议的当天，人民网和民族画报刊登了“中国各民族，为何如此团结”一文，用西北民族大学科研团队工作画面，报导了国家培训通用语人才成就。

本章编写人员：

2.1：吐尔地托合提、哈妮克孜伊拉洪

2.2：阿拉坦巴根那、娜仁高娃、飞龙、高光来、张晖、苏向东、刘瑞、王炜华、王斯日古楞、诺明花、哈斯、金罡、布音其其格、林民

2.3：于洪志、万福成

第 3 章 东盟语言信息处理

3.1 东盟国家语言概况

东盟包括 10 个国家：印度尼西亚(简称印尼)、马来西亚(简称马来)、新加坡、文莱、菲律宾、越南、泰国、老挝、柬埔寨和缅甸。由于历史原因，不同国家的语言政策有所不同，各国经济发展状况不同，信息化水平也不同，导致相应的自然语言处理技术的发展水平也不同。表 3-1 展示了各国语言使用的概貌。

表 3-1 东盟语言概况

序号	国家	语种数量（种）	官方语言	人口（万）
1	印度尼西亚	805	印尼语	27020
2	马来西亚	137	国语是马来语 官方工作语言是英语	3269
3	新加坡	20+	马来语、英语、华语、泰米尔语	545
4	文莱	10	马来语	43.7
5	菲律宾	186	菲律宾语、英语	11019.8
6	越南	100+	越南语	9758
7	泰国	72	泰语	6506
8	老挝	49	老挝语	727.5
9	柬埔寨	23	高棉语(即柬埔寨语)	1671.8
10	缅甸	100	缅语	5562

从语言形态看，印尼语、菲律宾语、马来语属黏着语，缅甸语、泰语、越南语、老挝语和柬埔寨语属孤立语，也称分析语。

从语系来看，印尼语、马来语、菲律宾语属南岛语系印度尼西亚语族；泰语属汉藏语系壮侗语族壮傣语支；越南语，一种意见认为属汉藏语系，另一种意见认为是南亚语系孟-高棉语族(该观点更流行)，还有一种认为属南亚语系越芒语族越语支(越南官方观点)；老挝语属汉藏语系侗台(壮侗)语族台(壮泰)语支；柬埔寨语又称高棉语，属南亚语系孟-高棉语族；缅甸语属汉藏语系藏缅语族缅语支。

3.1.1 印度尼西亚语言状况

印度尼西亚是一个多民族、多语言、多元文化国家，主要民族有

爪哇族(占 40.05%)，巽他族(15.5%)，马都拉族等。印度尼西亚语(简称“印尼语”)是印尼的国语，也是官方语言。印尼语和马来语属于同源语言，主要来源于古马来语，但马来语没有复杂的等级之分，词汇也没有复杂的形态变化，容易学习，属于开放式语言，能迅速吸收外来词汇和方言词汇。

印尼在 1945 年独立后，历届政府都将印尼语作为教育、政府和商业的基本语言大力推行。印尼语言政策导向主要由印尼教育与文化部语言建设和发展中心完成。1972 年印尼与马来西亚签订文化协议，共建两国语言工作者组成的“印度尼西亚-马来西亚语言大会”，1985 年文莱达鲁萨兰国加入该组织，更名为“文莱达鲁萨兰国-印度尼西亚-马来西亚语言大会”(Majilis Bahasa Brunei Darusslam – Indonesia-Malaysia)，新加坡是观察员国。该机构成立的目的是发展和培育马来语/印尼语成为高度文明的语言。

印尼政府也在积极提倡民众保护和传承各地(各族)语言和文化，对民族多样性和民族语言采取更为宽容和重视的态度。随着印尼经济的发展，印尼政府大力对外推广印尼语，尤其 21 世纪以来，每年向国外派出印尼语教师，出版对外印尼语教材，为外国留学生举办用印尼语讲故事比赛、写作比赛、提供奖学金等活动，设立印尼语言水平测试，力求通过印尼语发挥其在国际上的影响力。

3.1.2 马来西亚语言状况

马来西亚也是一个多民族、多语言、多宗教、多元文化的国家，全国有 30 多个民族，主要有马来人、华人和印度人，使用的语言主要有马来语、英语、华语和泰米尔语。马来语是国语，且是唯一的官方语言。英国曾殖民马来西亚近 200 年，在政治、经济、教育上已广泛使用英语。华语是马来西亚华人广泛使用的语言，包括普通话、闽南语、粤语、客家话、海南话、潮州话等。马来西亚印度人使用的语言主要有泰米尔语、孟加拉语、古吉拉特语、僧伽罗语等。分布于东

马和西马的土著马来人使用他们自己的语言，如特米亚语、雅贡语、马兰诺、爪威语等。

马来西亚有四种教学语言：马来语、华语、泰米尔语和英语。公立小学因教学语言不同分为国民小学和国民型小学。国民小学的教学语言主要是马来语，国民型小学分为华文小学和泰米尔小学，华文小学的教学语言是华语，马来语和英语是必修课。泰米尔小学的教学语言是泰米尔语，马来语是必修课之一。国民中学的教学语言是马来语和英语，华文独立中学的教学语言是华语。大中专院校的教学语言是马来语和英语。

媒体方面，马来西亚约有 50 种报纸，用 8 种文字出版；广播电台有 6 个广播网，用马来语、英语、华语和泰米尔语广播；有 2 个公共无线电视频道，无线电视 TV1 使用马来语播放节目，TV2 使用马来语、英语、华语和泰米尔语为观众服务。

马来西亚不遗余力推广马来语。宪法第 152 条明确规定马来语为国语。1967 年国会通过《国语法案》、后来的《1996 年教育法令》、1980 年 3M 制新课程、1985 年教育部实行的“综合学校计划”、1995 年政府提出的“宏愿学校计划”、2007 年教育部公布的《2006-2010 年教育发展大蓝图》、以及 2013 年公布的《2013-2025 国家教育发展大蓝图》都有明确加强国语学习和教学的内容。马来西亚语言政策的特点是强化国语，限制少数民族语言的使用和学习，并突显英语的地位。[77]

3.1.3 新加坡语言状况

新加坡的官方语言有 4 种，其中华语、马来语和泰米尔语是三大主族群的母语，英语是外来语，是新加坡的第一官方语言。

1965 年，新加坡独立后，“英语+民族母语实施双语制教育”开始从小学起正式实施，中学毕业时，要求每人至少掌握两种语言。新加坡的 4 种官方语言是报纸、广播和电视的媒介语言。在公共领域方

面，英语是社会通用的工作语言，是新加坡人的共同语；英语广泛应用于政府机关、行政管理和司法部门；政府的重要文件有英语版、华语版、马来语版和泰米尔语版；英语也是国际贸易、金融以及现代科学技术使用的语言。马来语和泰米尔语通常在家庭、族群内部使用。华语方言则在宗乡会馆、日常购物场所中使用。

在语言立法方面，1963 年新加坡作为马来西亚联邦的一个州时，州政府已经把马来语、华语、泰米尔语和英语定为新加坡的官方语言，并把马来语定为国语。1965 年独立后修订原有州宪法，保留了有关国语和官方语言的政策，同时规定：国会议员发言时，可以从 4 种语言中任选一种；国歌和军队号令使用马来语；1987 年，英语被官方正式采纳为第一教育语言。新加坡的语言政策与国家的政治和种族问题息息相关；语言政策讲求实用，服务于国家的经济发展；语言政策随着历史的变迁而适时调整，体现政策的连贯性和稳定性。[77]

3.1.4 文莱语言状况

文莱的国语是马来语，官方语言是马来语和英语，华语使用较广泛。马来语包括官方语言马来语(Bahasa Melayu，又被称为标准马来语)和文莱马来语(Brunei Melayu)等。文莱马来语常用于马来人之间非正式交往，是很有效的通用语。英语在文莱属于通用语言。文莱两种教学语言是英语和马来语。此外还有 10 余种少数民族语言，包括克达岩马来语、都东语(Tudong)、马来奕语(Belait)、杜松语、比赛亚语(Bisaya)等。

1984 年文莱宣布独立后，就实行马来语-英语的双语教育体制。从小学、初中、高中到大学，文莱把马来语的教育放在一个重要的位置。在双语制教育过程中，马来语授课的课程逐渐减少，使用英语授课的课程逐渐增多。司法领域是英语使用的另一个官方领域。[78]

3.1.5 菲律宾语言状况

菲律宾国内共 100 多种方言，分为本地语和外来语两大类，主要

有菲律宾语、英语、西班牙语、华语、阿拉伯语和各地方言。四大主体民族分别使用自己的语言，即比萨扬语(Isayang)、他加禄语(Tagalog)、伊洛克语(Ilocano)和比科尔语(Bicol)，这些语言还有各自的方言。目前90%的菲律宾人使用10种最主要的方言，如他加禄语、宿务语等。1959年，他加禄语更名为菲律宾语(Philipino)，1987年宪法确定菲律宾语为国语，是国家的第一官方语言。外来语中的西班牙语和英语也是官方语言。

1974年开始，菲律宾实行双语教学，公立和私立小学一、二年级的识字语言为本族语，同时学习英语和菲律宾语，从三年级开始教学语言为英语，少数族群的教学语言从五年级改为英语，菲律宾语作为一门课程从小学教授至大学。高校的人文科学、实用工艺、医疗卫生、品德教育、体育等科目必须用菲律宾语教学，数学、科学、音乐、艺术等科目使用英语教学，各地方言是辅助性的教学语言。进入21世纪后，菲律宾开始把加强“基于母语的教学”放在首位，各民族的母语作为教学语言使用，同时也加强菲律宾语、英语和西班牙语等语言的教学。

菲律宾的媒体以英语、菲律宾语和华语为主。广播电台节目中本地方言比菲律宾语使用更多。有线电视台节目大多用英语制作。英语是政府的正式用语，在外交、文化交流、科技、国际贸易等场合中广泛使用。英语也是学术界、文学创作、国家公职人员考试时使用的语言。在家庭、社区和非正式的商业活动中，英语和菲律宾语混合使用。菲律宾的语言政策政治化倾向明显，本土语言的象征意义大于实际价值。对少数民族语言权力的关注只停留在政策的层面，未真正落实到实处。[77]

3.1.6 越南语言状况

2013年，越南新宪法明确规定越语是国语、官方语言、通用语言和主要民族语言，越南使用过的主要文字有3种：第一种是公元前

2 世纪从中国传入的汉字，第二种是公元 11 世纪在汉字的基础上创制的越南人自己的文字—喃字(Chu nôm)，第三种是西方传教士等人发明的用拉丁字母拼写的国语字(Quôc ngu)。第三种国语字在 20 世纪成为越南全面使用的主导文字。相比汉字和喃字，拉丁化越语拼音文字更便于学习，为普及教育起了重要作用。

越南是多民族国家，有 53 个少数民族，使用的语言有 100 多种。这些语言都属于孤立语，有相同语音和语法特征。越南的教学语言主要是越语，10 多种少数民族语言被用于少数民族地区的学前和小学教育阶段。国家大力发展英语教育，英语在部分中学和大学成为教学媒介语。1945 年越南民主共和国诞生后，越语取代法语，成为各种社会活动的最主要语言。报纸和广播电台除了使用越语外，也出版少数民族文字的报纸，开通了少数民族语言频道。越南之声广播电台对外广播的语种有汉语、法语、英语、俄语、西班牙语、日语、泰语、印尼语、马来语、老挝语、柬埔寨语等。

越南语言政策的特点是国家保护、发展、推广国语和官方语言，尊重和支持少数民族语言和文化，重视外语教育，尤其是英语教育。[77]

3.1.7 泰语语言状况

标准泰语(简称“泰语”)是泰国的国语和官方语言。对于少数上层阶层的人群来说，标准泰语是第一语言，对于大多数泰国人来说，标准泰语是通过学校教育才习得。标准泰语是泰国所有教育阶段的教学和识字语言，外语教学中必须使用泰语作为教学语言。在媒体和公共领域，泰语被广泛应用于教育、行政、法律、宗教和大众传媒中。

语言政策方面，泰语于 1997 年颁布的《教育法案》第 26 条和第 27 条规定，所有教育材料必须适合泰国文化和身份，教科书中不应出现背离泰国文化的内容。2010 年，时任泰国总理阿披实·维乍集瓦批准泰国皇家学会(The Royal Institute of Thailand)起草的一项新的

国家语言政策，该政策主要涉及针对泰国人民和学生的泰语政策、针对泰国傣语系以及其他语系等本地方言的政策、针对经济用语、邻邦国家的语言和职业语言的政策、针对在泰国就业和工作的外国人的政策、针对盲人和听障人士的语言政策以及对于语言翻译、口译和手语翻译的政策。政策重申了泰语作为国家语言的地位，同时也加强了对英语、汉语以及泰国邻国语言的重视。泰国语言政策的特点如下：泰国政府对国内语言生活的管理是通过隐性的方式进行，没有关于语言政策的法律，只有 2007 年的宪法在人权部分提及民族语言权利；受民族主义影响严重，政策的出台都是基于国家安全和民族融合目的；没有明确的关于少数民族的语言政策；教育是泰国语言政策制定和实施的主要领域。[77]

3.1.8 老挝语言状况

1975 年老挝人民民主共和国(简称老挝)成立，将老龙族语言和改革后的老龙族文字规定为老挝人民民主共和国的普通话(标准老挝语)和官方文字。老挝语作为民族语言和官方语言在老挝全国推广。老挝是多民族、多语言国家，所用语言纷繁复杂。老挝政府推广采用标准老挝语的单一语言政策，有利于各民族的交流融合，但没有对众多的少数民族语言给予足够的重视。老挝经济极为落后，资金短缺，尤其是教育资金短缺，导致标准老挝语的推广受到限制。在推广过程中，各种外语仍然存在：法语曾是官方语言；英语是老挝对外的主要交际用语；俄语和越南语一度受到重视，20 世纪 80 年代后逐渐冷落；汉语在老挝的发展中相对平稳；由于地缘、亲情、史缘等关系，泰语对老挝人生活影响最深。[78]

3.1.9 柬埔寨语言状况

柬埔寨是多民族、多语言国家。高棉语是柬埔寨的国语，也是官方语言，在全国范围内广泛使用，英语和法语也是主要的政府部门工作语言，主要的移民语言有中国汉语普通话、闽南语和粤语。由于地

理位置和历史原因，泰语和越南语在柬埔寨使用率也较高，是主要的少数民族语言。

高棉语是全国中小学的法定教学媒介语。在高等教育阶段，高等学校有义务使用柬埔寨语进行教学，但用外语作为教学媒介语是寻求国际援助的重要手段。许多大学使用英语和法语课本。媒体用语主要是高棉语，也使用英语和中文。广播用语除高棉语外，还有法语、英语、汉语、越南语、老挝语和泰语。柬埔寨的公共交际语言有高棉语、英语、汉语等。

1953 年柬埔寨独立，但直至 1993 年，《柬埔寨王国宪法》第五款明确规定柬埔寨王国的国家官方语言和文字为高棉语，在全国范围内通用。在保护和发展国语同时，也鼓励外语学习和发展。其语言政策的演变和发展与殖民史密不可分，深受外来语影响。[77]

3.1.10 缅甸语言状况

缅甸是一个多民族、多语言国家，缅族作为主体民族占总人口约 65%。各少数民族都有自己的语言，其中克钦族、克伦族、掸族和孟族有文字。缅语是全国通用语、官方语言和公立学校的教学语言。

19 世纪末，英国开始对缅甸进行殖民侵略，期间曾把英语定为缅甸官方用语，强制英语为缅甸教学用语。1941 年日本取代英国开始对缅甸实行殖民统治，规定用日语代替英语教学。1948 年缅甸宣告独立，吴努政权开展提升缅语的一系列活动：以宪法形式确立缅语为官方语言，把缅语作为各级公立学校的教学语言；在全国推广缅语；在全国出版缅语书籍。1988 年军政府上台，加上英语全球化的进程，以及经济发展和国际合作的不断深入，英语重新受到重视。小学和初中阶段，缅语是唯一教学语言，但学生要学习缅语和英语两门必修课，高中阶段，用英语教授化学、物理、生物等科学课程，英语也是大学的教学语言。缅甸正规学习没有实行缅语-少数民族语言的双语教学，使少数民族语言的使用和传承受到限制。[78]

3.2 东盟官方语言信息处理综述

3.2.1 印尼语、马来语

1. 词法分析

词法分析是研究单词由较小单位即语素构成的方式。主要包括词干提取、词形还原、词性标注等任务，在印尼语、马来语词干提取和词性标注方面的进展较大，而词形还原方面的研究成果较少。

词干提取是指通过分离附加在单词上的词缀来提取单词词干的过程，印尼语、马来语通过将词缀按照一定规则附加在词干上派生出新词。但由于印尼语的发展自由性和快速性，词干规则不是一成不变的，基于规则的词干提取算法难以适应语言的快速演变。

Pisceldo 等人**错误!未找到引用源。**[79]将词汇分成动词、名词、形容词及其他四个类别，利用两层形态学(two-level morphology)构建了有限状态自动机，能够对多种两层形态学规则进行编码。基于此，Larasati 等人[80]增加了数词、副词等词汇类型来描述印尼语的两层形态现象，并实现了印尼语词法分析工具 MorphInd。对于给定输入，分析器会输出所有的语素和语素对应的标签。

马来语在词干提取方面研究成果不多。Sodhy[81]通过构建神经网络模型自动抽取马来语词语前缀和后缀。但是该方法在预训练时如果得到的前缀类型是错误的，在后续训练时得到“假”的输入，则导致最终学习到的转换前缀也是“假”的，而在词干抽取完后也缺少词干正确与否的验证过程。

由于马来语形态学现象大多数是级联的，将词素视为 n-gram 模型的处理单元是合理的。因此对于给定的马来语词素序列$w = m_1 m_2 \cdots m_k$，Sulaiman 等人[82]利用期望最大化算法 EM (Expectation-Maximization algorithm)训练 bigram 语言模型，用于求解马来语单词最大可能性的词素组合。该方法实现了自动学习马来语词缀规则。该算法偏向于较短的语素序列，无法处理具有较长语素序列

的情况词性标注。

词性(Part-of-Speech，POS)是词汇最基本的语法属性，进行词性标注便于判定每个词的语法范畴。词性标注是自然语言处理中一项非常重要的基础性工作，为词义消歧、句法分析、命名实体识别、问答系统、机器翻译等任务打下基础。

印尼语、马来语词法分析主要针对形态学变化规律，提出基于规则、统计及融合方法。也有少数研究通过深度学习方法来实现词干提取和词性标注。相比英语等同以拉丁字母为书写系统的通用语言，印尼语、马来语的词法分析研究进展不大，没有统一的高质量语料及统一的评价和测试标准，难以客观地比较各种方法之间的性能差异。

2. 命名实体识别

命名实体识别(Name Entity Recognition, NER)是文本中承载信息的重要单位。近年来，基于深度神经网络的 NER 方法受到了广泛关注。但基于深度神经网络的印尼语、马来语命名实体识别相关成果较少。

Gunawan 等人[83]构建了包含 4,139 个句子，合计 81,173 个词的命名实体标注集，并构建了一个基于 Bi-LSTM-CNNs 的命名实体识别模型，其 F1 值达到 77.47%。

Wibawa 等人[84]利用多个机器学习算法集成的策略来进行印尼语命名实体识别。

Alfred 等人[85]针对马来语提取了一系列语言规则，使用基于规则的方法来进行马来语的命名实体识别，其 F1 值达到 89.47%。

Sukmana 等(2021)研究了在印度尼西亚进行的名为 KGZ 的 Zakat 领域的知识图谱。它旨在提供有关 Zakat 和管理印度尼西亚 Zakat 的基本知识。KGZ 存在一些问题，首先，现有的印尼语命名实体识别(NER)是非限制性的，并且基于通用目的，其数据是从新闻等一般来源获得的。其次，Zakat 域中没有 NER 的数据集。Fu 等(2021)构建

了一个包含超过5万个句子(超过67万个标记)的印尼语NER 数据集(IDNER)，以缓解印尼语标记资源的短缺。在马来语中，相关的NER资源有限。因此 Fu 等(2021)提出了一个基于同源语言和迭代优化的标记数据集构建框架，构建一个包含28991个句子的马来语NER数据集(MYNER)。为了更好地整合NER的边界信息，他们还提出了一个多任务(MT)模型，该模型引入了一个辅助任务边界检测，以显式和隐式方式改进NER训练。此外，他们还提出了一种门控忽略机制，用于进行条件标签转移，减轻辅助任务的错误传播。实验结果表明，该模型在MYNER上获得了与基线相当的结果。

3. 句法分析

句法分析是指确定句子的语法结构或依存关系的任务。和英语、汉语相比，印尼语、马来语在名词短语和动词短语方面具有相对自由的词序[86]。因此由于词法多样性，印尼语、马来语句法分析更容易出现多个句法树及语法结构歧义的情况。

针对这种情况，学者在印尼语句法分析方面展开了积极的研究。Gusmita和Manurung[87]利用现有句法分析器symbolic parser[88]对训练集中的未经标注的印尼语句子进行分析，以得到对应的解析树，在此基础上通过构建概率上下文无关文法(Probabilistic Context Free Grammar, PCFG)模型来尝试解决结构歧义问题。但是该方法难以准确捕捉动词的子类别信息，导致构建的句法树不完整不正确。

Irmawati等人[86]考虑印尼语的形态特点，如词缀、省略、非谓语从句等等，在原有的 Stanford 依存标注体系基础上进行调整、拓展，设计了更适合印尼语的依存标注体系[89]，以解决在分配头部和关系时的歧义问题。Herlim 和 Purwarianti[90]提出一种移进-规约成分印尼语句法分析方法，在新的包括11,356和4,457个句子的INACL树库上，F1值达50.3%。

Rahman和Purwarianti[91]认为要提高印尼语句法分析器的效率，

应解决两个问题：(1)用于训练的树库句子缺乏多样性；(2)如何对各种句法分析技术进行有效结合。为解决这两个问题，他们构建了包含2,098 个句子的依存树库，并使用集成学习以最大化利用现有依存分析器。实验表明，基于图的算法性能要优于基于转换的算法，而集成学习并没有带来性能的显著改善。

与印尼语句法分析研究相比，马来语句法分析研究进展较为缓慢。Abidin 等人[92]使用自上而下的方法，结合马来语词典和语法规则，较早研究了马来语的句法分析。他们构建的句法分析器可画出正确句子的句法结构树，也可识别句子是否合乎语法。Noor 和 Jamaludin[93]更进一步利用句法分析的结果对不合乎语法的句子进行修正。他们利用上下文无关文法(context free grammar, CFG)对句子进行核实和匹配。符合 CFG 的将会可视化其句法树，不符合的会给出相关的修改建议。Hiloh 等人[94]指出马来语的语法是上下文无关语法，因此采用概率上下文无关文法来选择概率最大更精确的解析树。他们从多种来源收集简单马来语句子作为训练集。基于该训练集，他们构建了由词汇、语法规则及其概率组成的马来语的统计词汇语料库，采用 Cocke-Younger-Kasami(CYK)算法实现自下而上的句法成分解析，以帮助用户识别适当的解析树并解决马来语中的歧义问题。

以上工作表明，虽然印尼语、马来语的句法分析已有一定基础，但相比英语、汉语等通用语言还是相对滞后。大部分研究是基于研究者自身构建的小规模语料库，所以实验结果并不理想。但有些研究能够结合语法错误识别与纠正，能为高质量语料库构建提供帮助，从而提高句法分析的性能。

4. 语义分析

语义分析旨在解释自然语言句子或篇章各部分的含义。语义分析目前面临的问题是自然语言句子中大量存在歧义。学者对印尼语、马来语语义分析进行了初步的探讨和研究，主要集中在语义知识库构建、

词义消歧。相比词法分析及句法分析，语义分析研究成果较少。

Noor 等人[95]使用多个词典资源，如法-英-马词典、马来语词典等，初步构建了包含马来语和印尼语的 WordNet。该 WordNet 包含 49,668 个同义词集、145,696 个义项以及 64,431 个词汇。Mahendra 等人[96]利用英印平行语料和两个语言的 WordNet 自动构建了一批训练语料，再利用这批语料训练词义消歧模型。另外，他们还对比了分类器的选择、特征的选取、词干提取及停用词移除等因素对模型性能的影响。

印尼语、马来语语义分析研究刚起步，处于语义知识库构建阶段。而对于处于语言政策环境宽松的印尼语、马来语，其词汇语义歧义现象更为频繁，语义分析至关重要。

5. 语料库与机器翻译

印尼语、马来语属于低资源语言，相关机器翻译的语料如英印、英马平行句对较为缺乏，成为研究机器翻译的一大阻碍。主要研究成果是基于统计的机器翻译，也有学者利用深度学习实现神经机器翻译。

印度尼西亚技术评估与应用署(Badan Pengkajian dan PenerapanTeknologi, BPTT) [97]构建了较大规模的英印单语语料和双语语料，并利用 SRILM、Giza++以及 Moses 开发了一个双向的英-印统计机器翻译系统。Hermanto 等人[98]初步训练了一个统计翻译模型和一个基于 RNN 的神经翻译模型，实验表明神经网络机器翻译的效果优于统计机器翻译。

Yeong 等人[99]提出一种将 SMT 和 NMT 结合起来进行英语到马来语翻译的混合方法，并使用 LSTM 架构，混合 MT 在计算机科学领域和新闻领域的 BLEU 分数从 21.21 和 48.35 分别增加到 35.97 和 61.81。Wang 等人[100]指出由于印尼语-英语的双语文本较少，而马来语-英语的双语文本相对较丰富，且印尼语和马来语有极大的相似性，因此他们用了三种方法将马来语文本转换为印尼语文本，以构造规模相当的印-英双语文本，再用于统计翻译模型的训练。Octoviani 等人

[101]提出利用循环神经网络(Recurrent Neural Network, RNN)和注释分离(Annotated Disjunct)实现基于短语的英语-印尼语机器翻译的方法。在 70 个英语短语上测试，取得的准确率为 88.57%。Yusoff 等人[102]认为现有的大多数马来语-英语机器翻译都是逐词翻译，没有解决歧义问题，因此他们提出了一个基于语义消歧的机器翻译方法。Yeong 等人[103]利用词典和词形还原工具解决翻译过程中的未登录词问题，以提升英语-马来语机器翻译系统的性能。实验表明，该方法将 BLEU 分数提高了 2.51。

郑锫涛等人[104]针对平行语料的对齐，提出了改进的段落对齐与句对齐方法，在结合锚点和词典实现段落对齐的基础上，采用基于置信区间的长度模型进行句子对齐，从而构建了中印平行语料库。

Qiu 等人[105]提出了一种印尼语-汉语双语词典构建的研究，使用单语词汇嵌入和双语种子词汇来构建共享双语词汇嵌入空间。他们进一步探讨了不同单语的语言特征对种子词典选择及模型性能的影响，发现虽然使用单语的一个语言特征训练的模型不如集成四种语言特征的效果好，但是学习印尼语和汉语的共有语言特征明显改进了单词互译的效果，因此表明语言特征在区分共享单词嵌入空间的语义边界方面具有进一步研究的价值。

6. 拼写检查

拼写检查旨在检索文本输入中因人为拼写错误导致的文本错误。由于印尼语、马来语文本中存在较多拼写错误，该现象吸引了学者的关注，并开始研究拼写检查。

Soleh 和 Purwarianti[106]研究针对印尼语非词错误的拼写检查器的构建，使用词法分析器和词典来实现错误识别，并基于编辑距离和最优子序列的方法来进行错误修正。

Irmawati 等人[107]提出一种自动生成含介词错误的印尼语训练数据的方法。Fahda 和 Purwarianti[108]提出一个融合规则和统计方法的印

尼语文本检查器。他们的检查器分为规则匹配、拼写检查和语法检查三个模块。规则匹配包含 38 条规则，可侦测、修正和解释在标点符号、词语选择和拼写方面常出现的错误。

Mawardi 等人[109]研究基于编辑距离利用有限状态自动机和 N-gram 实现印尼语文本的拼写校正系统。实验表明，bigram 相较于 unigram 和 trigram 具有最高的校正命中率。

Lin 等人[110]将语法错误纠正(Grammatical Error Correction, GEC)视为一项多分类任务，并提出印尼语 GEC 的框架。该框架集成了不同的语言嵌入模型和深度学习模型，以纠正印尼语文本中的 10 种词性的语法错误。此外，他们还构建了一个印尼语语料库，用作印尼语 GEC 研究的评估数据集。实验表明，基于字嵌入的长短期记忆(LSTM)模型的整体性能达到了最好的结果，平均 F1 为 0.477。

相比之下，在马来语拼写检查方面成果较少。Kasbon 等人[111]开发了一个马来语句子检查器。用户测试表明该系统更适合在小学和初中推广使用。而 Basri 等人[112]设计了一个针对马来语博客的拼写检查器，他们利用编辑距离对候选词排序。Noor 和 Jamaludin[113]设计了 BMTutor，旨在促进人们对马来句子的结构和语法的学习。该系统使用上下文无关文法对句子进行句法分析，并在句法分析基础上对句法树进行可视化，同时可检查句子中存在的错误并给出相应的修改建议。

7. 情感分析

随着互联网技术的普及，社交平台上产生了大量用户参与的、对于诸如人物、事件、产品等有价值的评论信息。潜在用户会通过浏览这些主观色彩的评论来了解大众舆论对于某一事件或产品的看法。这种现象吸引了广大学者关注，印尼语、马来语情感分析研究成果逐渐增多。

Franky 等人[114]翻译了现有的 4 个英语情感词典，得到对应的 4 个印尼语词典，再对这 4 个词典进行交与并的操作，得到最终的印尼

语主观性词典。Koto 和 Rahmaningtyas[115]也构建了一个印尼语情感词典 Inset。他们从推文中获取相关词汇，然后通过人工对每个词根据其情感倾向程度进行打分，并利用词形还原和同义词集对原词集进行拓展。

Lunando 和 Purwarianti[116]在对印尼语推文进行情感分类的同时，还探讨文本中的讽刺现象。他们在原来的情感分类模型上，加入了否定信息及感叹词的数目两个额外特征，以帮助侦测讽刺现象。

Fauzi[117]探讨随机森林算法在印尼语情感分类中的作用。他们考虑利用词频 TF，逆文档频率 IDF 以及 TF-IDF 三种加权方法确定词袋(BOW)特征的特征向量值。实验表明，随机森林用于情感分类的性能良好，平均 OOB 评分为 0.829。

以往的情感分类多注重对特征的选取和构建，而 Sadanandan 等人[118]不把重心放在特征选择，而是采用知识库和机器学习结合的方法来改善马来语情感分类的性能。他们利用 1,861 个人工标注的句子进行十折交叉验证，结合知识库和机器学习方法的总体准确率为 94.34%。而 Al-saffar 等人[119]提出一种基于语义定位和机器学习方法的马来语情感分析分类模型。他们使用三个单独的情感分类器和集成分类器来评估分类准确性。通过对马来语评论语料库(MRC)进行广泛的对比实验，证明特征提取提高了基于集成模型的马来语情感分类器的性能。但是，结果取决于三个因素：特征、特征数量和分类方法。

8. 自动文档摘要

自动文档摘要是自动地将文本或文本集合转换成简短摘要的信息压缩技术。目前印尼语、马来语文本自动摘要的研究成果还不多。

使用 TF-IDF、预定义的文本特征等对印尼语新闻文章进行排序与摘要抽取。使用 LDA 算法与遗传算法相结合，基于贝叶斯来提取印尼语新闻文章的摘要。Koto 等人[120]构建了一个公开的摘要数据集，即聊天数据集及其摘要，其中包含抽取式和生成式版本。这项工作是

印尼语摘要抽取工作标准化研究中的一大进步。Kurniawan 和 Louvan[121]构建了一个公开的印尼语自动摘要数据集 INDOSUM，并在该数据集中完成了几个模型的实验，其中 NEURALSUM 方法取得了最好的结果。Cai 等人[122]构建了印尼语自动摘要数据集，并提出了一种基于句子相似性聚类的自动文本摘要方法。实验表明，此方法可以确保摘要的完整性、关键性和重要性，并且减少摘要的信息冗余。在评估中，他们的方法取得了良好的效果，在 ROUGE-1，ROUGE-2，ROUGE-3 上的$F1$ 得分超过了所有基线模型。

Puspitaningrum(2021)调查了近期几种抽象的摘要方法：T5，Pegasus 和 ProphetNet，研究了预训练模型对英语和印尼语的几个维基百科数据集的影响，并将结果与维基百科系统的摘要进行比较。T5-Large，Pegasus-XSum 和 ProphetNet-CNNDM 提供了最好的总结。影响 ROUGE 性能的最重要因素是覆盖范围、密度和压缩率。

9. 未来挑战和发展趋势

印尼语、马来语自然语言处理研究成果分布比较宽泛，涵盖了词干提取、词性标注、句法分析、语义分析等底层技术及机器翻译、拼写检查、情感分析、命名实体识别等上层应用技术。印尼语、马来语词法分析、机器翻译、情感分析等方面取得了较好的进展，但语义分析、命名实体识别、拼写检查等方面研究进展不大。同时针对这两种语言的基础资源、开放数据平台及开源的语言处理工具也较为缺乏，也少有成熟可用的文本分析系统。这在一定程度上限制了印尼语、马来语自然语言处理技术的广泛研究。

在资源开放方面，印尼语、马来语仅有少量公开的平行语料、依存树库：School of Educational Studies 提供了超过 4,800 马来语词条，印度尼西亚大学计算机科学学院研究团队基于 PAN Localization 提供的英印平行句对人工审核及短语句法结构标注的 1,000 句印尼语句子 Indonesian Tree bank、包含 1 万句已进行词性标注的印尼语句子

Indonesian POS Tag，词干提取工具、词性标注工具、句法分析工具、语义分析工具及词法分析工具。除此以外，其他领域几乎很少有公开的加工语料和工具供研究人员进一步研究。

目前针对印尼语、马来语的自然语言处理主要采用相对传统的技术和方法，由于语料规模不够，深度学习等前沿理论、技术和方法的应用效果不理想。这种现状也在一定程度上导致印尼语、马来语自然语言处理技术和其他语言如英语等通用语言相差较大。

尽管印尼语及马来语相关自然语言资源、处理方法和应用研究已经得到了一定进展，但是面对印度尼西亚及马来西亚国家语言政策，较为自由的语言文化发展氛围，其语言具有多样性和复杂性，在自然语言处理研究上仍有很大的探索和研究空间。在语言资源及共享平台建设方面，针对低资源语言的计算模型及语言处理技术是主要研究方向。

针对语言资源缺乏的印尼语、马来语自然语言处理研究，后续需进一步研究如何采用半自动或全自动的手段构建高质量、大规模的语料库。一个直观的想法是借助资源丰富的语言，利用语言相似性，通过跨语言映射的方式为目标语言获得可用的标注信息，从而自动构建待分析语言资源库。其次，通过对现有研究成果的归纳分析发现，印尼语、马来语并没有公认、完善的语法规范，例如在词性标注上不同学者采用自己设定的标注体系，若能制定一套系统成熟的语言规范，将有利于后续各个领域语料库的构建与完善。

由于印尼语、马来语自身的语言复杂性，一些适用于通用语种的研究方法并不能完全适用于这些语言的研究。因此针对印尼语、马来语的语言特点，在基础理论和计算模型上进行创新性研究，探讨适合印尼语、马来语的语言计算模型是今后研究的一个重要方向。在统一、规范化的词性标注体系或依存关系标注体系下和现有深度学习模型的基础上，结合印尼语、马来语的语言特性，提出创新、适用、高效

的方法。

3.2.2 菲律宾语

1. 词法分析

(1)词干提取

菲律宾语的词缀系统非常复杂，包含前缀、中缀、环缀、后缀、重复及以上多种词缀叠加。菲律宾语中的重复可以是单词部分重复或全部重复。多种词缀叠加是菲律宾语动词中常见的语言现象。例如单词 pinanglilibang-libang，通过词干 libang 附加前缀 pang，而前缀 pang 又叠加中缀 in 组成 pinang，并且部分重复 li 及全部重复 libang 来构成。由于菲律宾语构词的复杂性，形态分析成为其自然语言处理领域的基础任务。

由于菲律宾语的动词形态变化较其他词类更为丰富，所以形态分析研究主要针对动词。大部分研究都在利用菲律宾语的形态学变化规律的基础上，提出了基于规则的形态分析方法，其准确率较高。Roxas 和 Mula[123]设计了一个动词词法分析器，能给出该动词的基本形式，所含词缀及对应的时态(过去时、现在时及将来时)。Bonus[124]提出了一个不限于动词的基于词典的词形还原算法 TagSA，其中考虑了词缀、重复及复合等情况。在 6 千多个词语上进行了测试，取得了不错的效果。

(2)词性标注

与英语相比，菲律宾语同样具有后缀、大写字母等可用于确定 POS 的语言特征。除此以外，菲律宾语的词性标注离不开前缀、中缀、环缀、重复等有用的语言信息。

Erlyn 和 Yuji[125]探讨了影响菲律宾语词性标注效果的因素，考虑了菲律宾语单词的形态结构、形态信息（如词缀）作为训练 POS 模型的输入。实验使用了菲律宾德拉萨大学(DLSU)的人工标注数据，包括 114,096 词条，POS 标注集包括 9 个粗粒度标签、60 个特定标签、

5 个标点符号标签及其他符号的标签，准确率高达 93%以上。

Go 和 Nocon[126]构建了基于 Stanford 词性标注器的菲律宾语词性标注。核心算法为最大熵循环依赖网络，在设计特征时考虑到了词汇的形态及句子内部的语码转换信息。使用的词类标记集也是 MGNN 标记集，最终得到的准确率为 96%。

菲律宾语句子中单词顺序自由，导致菲律宾语不可以通过分析目标词前后词汇的分布概率来预测目标词的 POS 标签，将 POS 标注视为序列学习任务则无法很好地学到菲律宾语语法结构模式，从而导致实验效果不好；而标注语料的缺乏也限制了词性标注工作的开展。

2. 句法分析

菲律宾语句子中各组成成分顺序较为自由，不具有主谓一致的语法特点，且句子的焦点成为主题而不是主语。这些语言特征成为句法分析的一大障碍，导致适用于菲律宾语句法分析的算法较少，其研究成果也很少。

Clark[127]尝试利用词汇功能语法(Lexical Functional Grammar, LFG)作为计算模型来捕获菲律宾语的信息，实现了一个用于菲律宾语书面句子语法分析并输出句子功能结构的系统 FiSSAn，目前只能处理陈述句。

Alcantara 和 Borra[128]使用无监督的统计方法对菲律宾语句子进行构成成分的划分，在对句子进行词形还原和词性标注后，统计分析所有出现的词性标注序列，以生成划分构成成分的规则，由此得到的规则库即可划分后续句子的构成成分。

Manguilimotan 和 Matsumoto[129]首先研究了针对菲律宾语的依存句法分析，采用基于图的最大生成树算法，探索粗细粒度的词性、词根和形态等特征对句法分析模型性能的影响。在 2741 个句子上训练和测试，对于无标签依存关系，其平均准确率为 78%，而对于整个句子，其平均准确率仅为 24%。实验结果表明，当词性信息不够准确时，

加入形态信息有利于提高句法分析器的性能。

3. 语义分析

目前对菲律宾语语义分析主要集中在语义知识库构建、词义消歧。Mistica 和 Baldwin[130]实现了基于 CRF 的语义分析器，以识别菲律宾语中的谓词-论元结构。相比词法分析及句法分析研究，菲律宾语语义分析研究成果较少，而且其语义知识库构建仍处于初级阶段。

4. 机器翻译

菲律宾的机器翻译始于上世纪 90 年代后期，涉及菲律宾国家的两种官方语言：菲律宾语和英语。菲律宾语的机器翻译研究取得较大进展，其研究方法涵盖基于转换、基于语料库、基于统计和基于深度学习的方法。

Borra[131]探讨了词汇功能语法作为文法形式，发现功能结构(f-structure，f 结构)和组分结构(c-structure，c 结构)有助于识别翻译错误，并在此基础上提出了基于词汇功能语法的英菲机器翻译系统。基于转换的方法，翻译的效果受限于语料规模及转换规则，无法翻译词典外的词汇(Out of Vocabulary, OOV)。

基于语料库的机器翻译方法和传统基于规则的方法有很大不同，基于语料库的方法并不对目标语言进行深入复杂的语法分析，也不通过规则转换，而使用源语言和目标语言相对照的双语或多语语料库直接或间接地进行翻译。Roxas 等人[132][133]提出了基于转换规则和基于语料库混合的方法。Ong 等人[134]提出了一种基于模板的机器翻译系统，该系统从给定的双语语料库中提取模板，并常见词汇过滤及组块对齐算法来提高提取模板的质量。

基于统计的机器翻译方法是一种间接地使用语料库的机器翻译方法，它是通过双语句对的对齐，分析词汇共现的可能性来计算源语言的某一个词映射到目标语言的一个或多个(或零个)词的概率。Ang 等人[135]构建了一个基于 Moses 的菲英统计翻译系统 FEBSMT，所用

的实验数据来源于 22031 句旅游领域的英菲平行句对。该系统可接受用户反馈，并周期性汇总反馈数据以对系统做增量式训练，提升系统性能。

由于自动构建平行语料库方法的可用性，基于深度学习的菲律宾语机器翻译研究取得了一定进展。Tacorda 等人[136]利用 100,000 英菲平行句对训练 RNN 模型，并集成字节对编码(byte pair encoding, BPE)以减少 OOV 翻译错误。BPE 将一个词条分成可识别的字符序列。因此如果已通过 BPE 识别出训练数据的词干和词缀，则可以识别训练数据中不存在的词条。但是 BPE 无法处理误将词干的字符序列识别为词缀的情况。而针对 OOV 翻译的问题，Lazaro 等人[137]利用领域适应技术预处理训练数据，从而减少 OOV 的概率。

菲律宾语除了具有句子结构成分顺序自由的特点外，其动词拥有时态和焦点的特点及词缀包含前缀、中缀、后缀、环缀及重复等复杂的形态变化特点，这些都给菲律宾语机器翻译带来一定的挑战。由于菲律宾语目前还没有成熟可用的语言工具，如词干提取、词性标注等工具，菲律宾语机器翻译仍有很大的探索和研究空间。

5. 情感分析

Andrei[138]构建了一个小规模的面向 Twitter 的菲律宾语情感词典 LIWC(Linguistic Inquiry and Word Count)。Regalado 等人[139]研究了菲律宾语文本的主观性分类。他们以 TF-IDF 为主要特征，分别对文档和句子用 C4.5、朴素贝叶斯、KNN 和 SVM 等算法进行了主观性分类。在文档级别，SVM 取得最高的准确率为 95.06%；而在句子级别，朴素贝叶斯取得最高准确率为 58.75%。Pippin 等人[140]首先对菲律宾人发的推文进行情感分类，情感分类体系包含七个类别：开心、伤心、愤怒、惊恐、惊奇、厌恶及中性。然后使用朴素贝叶斯算法在 300,000 篇推文上测试，其准确率约为 70%。

Lapitan 等人[141]利用众包方式构建了一个小规模但高质量的

Twitter 情感语料库。他们的情感分类体系中包含九个类别：愤怒、期待、愉快、伤心、信任、惊奇、厌恶、恐惧及其它。在随机选取了778 篇菲律宾语推文和 570 篇英语推文后，依托 CrowdFlower 平台对这些推文按照指定规范进行人工标注，实验表明现有的语言资源和工具还不足以对推文进行准确的情感分类。

菲律宾语情感分析主要是有监督的，依赖人工标注情感分类。而情感分类体系因不同学者而异，且实验数据大多数是基于自己构建的小规模数据，因此无法客观地比较各种方法的效果。

6. 命名实体识别

菲律宾语命名实体识别研究成果较少。Eboña 等人[142]利用最大熵来实现菲律宾语小说摘录的命名实体识别。Alfonso 等人[143]提出利用条件随机场实现菲律宾语文本命名实体识别系统 NERF-CRF，其 F 值在 80%-83%，在地名和机构名的错误率分别为 2%-42%和 13.10%-33%。

7. 拼写检查

菲律宾语拼写检查研究除了基于规则的方法，有不少研究开始考虑综合其他自然语言处理工具来提高纠错准确率。

Dimalen 和 Dimalen[144]，Oco 和 Borra[145]，Go 等人[146]也实现了一个菲律宾语拼写检查器 Gramatika，一种基于词典及规则的拼写检查器，用于检查词语拼写错误、语法错误、漏词等情况。在不足 300 个带有错误的句子上测试准确率为 64%-83%。

由于菲律宾语语言资源及高效准确可用的语言分析工具的缺乏，与英语相比，其拼写检查研究严重滞后。Tsao 和 Wible[147]及 Huang 等人[148]通过实验表明 POS 的引入使得拼写检查性能显著性提升。

8. 语料库构建

在人工构建菲律宾语语言资源(例如词典、形态信息、语法规则库和语料库)方面有很大进展。除此以外，由于人工构建语料库的内

在困难，不少学者开始研究自动抽取高质量语言资源的技术。

Tiu 和 Roxas[149]提出了一种从可比语料中自动提取双语词典的方法，其中英语为源语言，菲律宾语为目标语言。他们结合上下文抽取、聚类技术，并使用词性标签来定义单词的不同含义。Dita 等人[150]初步通过人工构建菲律宾国家语言的在线语料库，包括菲律宾语、宿雾语、伊洛卡诺语、希利盖农语和菲律宾手语。前四种语言包含 250,000 单词的文本，而菲律宾手语包含 7,000 个视频。该在线语料库还提供了用于语言分析的自动化工具，例如字数统计。该项目后续考虑自动获取文本、语音、视频等多模态语料资源。

文献[151]的工作是为德拉萨大学(De La Salle University - Center for Language Technologies)研发了英菲机器翻译系统服务[152]。除此以外，面对有限的菲律宾语语言资源，基于菲律宾语语言委员会提供的词典，他们还构建了一个英菲词典，包含词条的形态学信息，如词性标签。

Borra 等人[153]讨论了菲律宾语 WordNet 的构建，探讨了菲律宾语的形态用于构建分析器和生成器以支持 WordNet 中的词干以及词缀序列对的收集。El-Kishky 等人[154]应用 URL 匹配规则从 commoncrawl 语料库中爬取高质量的跨语言文档数据集，包含 92 种不同语言(包含菲律宾语、印地语、德语等)与英语对齐的文档对。他们首先使用人工注释来直接评估该数据集的质量，而后通过评估下游任务，即利用该对齐语料训练的机器翻译模型评估数据集的质量。

Sagum 等人[155]提出基于决策树和 n-gram 模型的半监督方法来构建菲律宾语语义知识库 FilWordNet。在 500 篇文档中(包含 25,618 单词，其中 15,377 菲律宾语单语单词)上测试，正确提取词干且进行 POS 的准确率高达 86.29%。

Velasco 等人提出了一种仅使用未标记语料库和基于句子嵌入的语言模型从头开始构建 wordnet 的自动方法，生成一个新的

wordnet—FilWordNet，它取代并改进了过时的菲律宾 WordNet。

9. 未来挑战和发展趋势

总体来说，在菲律宾语自然语言处理领域，语言资源不足，特别与英语、汉语等语种的自然语言处理研究相比还存在较大差距。现有研究相对宽泛但不够深入，在词法分析、句法分析、语义分析等底层技术及机器翻译、情感分析、拼写检查等应用技术都有一些成果。其中，机器翻译的研究取得了较快的进展，而在句法分析、语义分析、命名实体识别等方面的研究成果相对较少。菲律宾语的机器翻译主要是英语-菲律宾语的翻译，很少涵盖其他语言。这与菲律宾国家的语言政策有关，因为菲律宾第二官方语言是英语，菲律宾政府和学术研究机构在英语和菲律宾语的语料构建及英菲机器翻译上投入较大的人力和物力。而菲律宾语与其他语言对照的平行语料缺乏，研究投入不足。

虽然菲律宾语在语料库自动构建方面的研究取得了一定进展，但是相较于英语、汉语等通用语种，菲律宾语仍属于语言资源较为缺乏的源语言。大部分语料库构建研究旨在收集英菲平行句对或词对，主要服务于机器翻译；在自然语言处理其他领域的语料资源构建研究非常少。由于深度学习算法高度依赖于高质量、大规模的标注语料，导致无法有效运用深度学习方法于词法分析、句法分析、命名实体识别等方面。

在信息大爆炸时代，信息的提取和精炼成为一个重要的研究课题，而文本自动摘要是解决信息爆炸问题的关键技术，跨语言自动摘要技术可以让人们快速了解不同国家和地区的信息。然而，根据文献调查发现，目前菲律宾语文本自动摘要研究几乎处于空白状态。

综合以上对菲律宾语自然语言处理现状可见，英语-菲律宾语平行语料较为丰富，有力推动了机器翻译的研究进展。面对丰富的英语-菲律宾语平行语料，如何通过跨语言处理技术，构建汉语-菲律宾语

平行语料库成为我国研究汉语-菲律宾语机器翻译、跨语言自动摘要等任务的首要解决问题。

针对菲律宾语的其他自然语言处理领域语料匮乏的问题，同时在词法分析、句法分析、语义分析等任务上无法使用海量无标注语料进行深度学习，因此构建相关领域较大规模、开放的标注数据显得十分必要。面对资源缺乏的基础问题，尽管菲律宾语形态变化丰富，但只要总结足够多的形态规则就可以构建形态学信息语料库；而正确的形态学信息可为词性标注和句法分析等提供重要语言特征，有利于提高其他自然语言处理任务的性能，从而利用半监督的资源构建技术促进其他领域语言资源构建。

在大规模、高质量、开放的语言资源构建的前提下，深度学习应用于菲律宾语自然语言处理方法研究成为可能。在基本理论和模型创新基础上，鉴于菲律宾语句子语法结构较为灵活，结合基于规则、基于统计和深度学习的方法，在一定程度上解决由菲律宾语复杂的语言特征造成的诸如词义多样、句法结构歧义等问题，从而推动命名实体识别、句法分析、语法纠错、知识图谱构建及语义分析等方面的研究。

最后，考虑到信息爆炸时代下文本自动摘要技术的重要性，借鉴其他语言的文本自动摘要研究技术，探讨基于规则、基于图模型、基于结构等方法对菲律宾语文本自动摘要的适用性，以填补菲律宾语自动文摘研究的空缺也是未来重要的研究方向。

3.2.3 越南语

1. 词法分析

从越南语的历史发展来看，越南语与朝鲜语、日语和琉球语一样自古受到汉字文化的深远影响。在近代，法国殖民者开始了一系列的去中国化运动，形式上较为接近西方语言，经过中西两种语言文化的影响，如今的越南语同时具有中西两类语言的特点。

(1)分词

首先从文字形式来看，越南语更加接近西方语言，如“Xungđột Nga-Ukraine gây trở ngại lớn chotiến trình toàn cầu hóa(俄乌冲突对全球化进程构成重大障碍)”，越南语使用拉丁字母组成词语，词语之间使用空格隔开，与西方语言十分相似，但是越南语却属于单音节语言，每个音节分别发音并且几乎每个音都有其所表达的涵义，这一特点就与中文的特点十分相似，例如：Cây(树)、Cỏ(草)、tay(手)、ăn(吃)、Điện(电)等，每一个音节都反映一个对象或概念。

越南语不止有单音节词语，同时也有复合词，就是由两个及以上的单音节词组合构成一个新的复合词，如：Trung Quốc(中国)、yêu vàquý(爱情)、Tin tức(新闻)等，所以越南语的语句与中文类似，是由词语构成而不是单个音节组成。

相比于英语文本可以借助空格来进行文本分词，如 China、school、people 等，每个连续字符串都有各自所表达的涵义，而如果要用越南语表达这三个词的涵义则为：Trung Quốc(中国)、Trường học(学校)、Mọi người(人民)，所以越南语就不能像英语一样按照空格来进行分词，否则会失去词语的涵义，所以在分词方面，越南语分词相比汉语更复杂。

最早的分词算法主要是基于词典的，包括最大匹配 (Maximum Matching, MM) 算法和逆向最大匹配 (Reverse Maximum Matching, RMM) 算法。基于词典的算法易于实现，但其性能很大程度上取决于词典的规模和质量。Liu et al.(2016)提出两个新的度量：平方重叠率(SOR)和松弛平方重叠率(RSOR)，用于解决字典大小如何影响越南语分词的性能的问题，并验证了它们的有效性，结果表明大小适合的字典能更好地实现基于字典的越南语分词的最佳结果。

机器学习方法也被应用于越南语分词的任务中，如一些研究使用支持向量机和条件随机场进行分词任务，并结合音节连词、字典等特征，准确率可以达到 94%(Nguyen et al.,2006)。VnTokenizer 是由越南

本国河内大学采用基于最大匹配和N-Gram模型开发的越南语分词工具，许多研究引用 VnTokenizer 将分词结果用于构建大型越南语语料库。歧义问题广泛分布在越南语句子中，影响分词的准确性， Xiong et al.(2016)提出了一种基于 CRF(条件随机场)和交叉歧义模型的越南语分词方法，结合越南语词汇特征，将越南语的基本特征融入条件随机场，从训练集中提取 5377 个歧义片段，选择统计特征、歧义场内部特征和歧义上下文特征，放入最大熵模型和交叉歧义模型中，然后纳入分词模型，经实验分词准确率达到 96.55%，准确率和召回率相比 VnTokenizer 提高了 1.34% 和 0.63%。

Zheng et al.(2022)[156]使用深度神经网络方法针对越南语分词的歧义问题，提出了一种基于改进的长短期记忆神经网络(LSTM)的越南语分词处理技术，由 LSTM 编码和 CNN 特征提取部分组成，将分词处理任务细化为分类问题和序列标注问题，可以自动获取分词字符和词级的有用特征，避免了局部上下文窗口大小的限制，将分词处理任务细化为分类问题和序列标注问题，在越南新闻网站爬虫数据集进行验证，该方法准确率达到 96.6%，召回率达到 95.2%，F1 值达到 96.3%。

(2)词性标注

侯中熙等(2022)将基于SVM的SVMTooL应用到越南语词性标注上。标注集按照越南语的词性和符号共分为 28 种标注，训练语料包含 25 万词，实现了的越南语词性注，取得了较好的效果，准确率达到 96.01%。Chen 等[157]提出了一种基于多类别词的消歧模型越南语词性标注的新方法，使用爬虫程序在越南新闻网站上获取语料库，包括经济、政治、文化和军事领域，然后对这些获得的语料进行处理，形成文本语料，并整理定义了 19 种标注集，完成了 27878 句人工标注，实验表明，所提出的方法可以有效地进行越南语词性标注，准确率达 95.22%。

与此同时，当前研究中也产出了一些越南语词性标注的分析工具包。Le-Hong et al.(2010)对最大熵方法在越南语文本的词性标注中的应用进行了实证研究，该方法用于标记越南语文本并给出 93.40%的整体准确率和 80.69%的未知单词准确率，并以此为基础开发了一个开源的名为 VnTagger 的词性标注软件包。另一种工具 JVnTextPro，可以像 VnTagger 工具一样标记越南语单词，它基于 CRFs 模型和最大混沌模型以及用于选择的上下文模型。JVnTextPro 使用来自 VietTreeBank 的文本数据库的 10,000 和 20,000 个句子的训练数据集进行训练。Vu 等[158]开发的 VnCoreNLP 工具包包含词性标注的功能，可以达到每秒标注 25K 个词汇并且具有 95.88%的准确率，表现优于 BiLSTM-CRF 模型。Quach 等[159]对较为常用的几个越南语词性标注的工具包进行了对比，包括 VnTagger、RDRPOSTagger(Java 版)、JvnTextPro、VNCoreNLP，经过实验对比，JvnTextPro 的准确率最高。JvnTextPro 和 RDRPOSTagger 工具总是有一个稳定的准确度水平，因为这两个工具处理单词的能力相当稳定，尤其是在文字处理部分，复合词。当存在双词时，上述两个工具的复合词会直接标注词，而不是拆分和丢失词的含义和性质，因此与其他两个工具即 VnCoreNLP 和 vnTagger 相比，准确率显着提高。

2. 句法分析

越南语是一种固定语序的语言，由固定的语序构成主谓宾(SVO)，也就是说，他们一般的语序为：主语＋谓语＋宾语。汉语、越南语在对所属关系的表示方面与英语不同，英语有明显的标志性词语用以界定，而汉语、越南语则没有特定分隔或标志性词语，因此存在着结构方面的歧义。越南语句法分析目的是构建一定规模的越南语短语树和依存树，同时增强越南语句法分析的准确率和效率。

Manh et al.(2008)利用基于转移的依存分析器 MaltParser 和基于图的依存分析器。MaltParser 解析器利用投票算法作为校正规则，为

越南语依存分析提供了良好的基线。李英(2017)提出一种融合越南语语法特征与改进 PCFG 的越南语短语树库的构建方法。该法能自动分析出越南语的短语结构树，解决了越南语短语树库的构建问题。他们首先通过分析越南语语法特征，制定了越南语语言特征集；然后利用 Inside-Outside 算法从人工标注的越南语短语树获取 PCFG 模型中的语法规则集，最后将越南语语言特征集作为语法规则集的补充融入 PCFG 模型，并利用得到的新模型完成越南语短语树库的构建。

Nguyen 等[160]将 LSTM easy-first 依存分析与预训练的词嵌入和字符级词嵌入相结合，在越南依存树(VnDT)上实现了 80.91% 的无标签标注得分和 72.98% 的有标签标注得分的准确性。Thi 等[161]研究了双向长短期记忆网络模型在越南语的基于转换和基于图的依赖解析中的使用。经过在其构造的语料库中实验，无标签标注得分为 84.45%，有标签标注得分为 78.56%。

3. 语义分析

相对于英文和中文来说，越南语浅层语义分析的相关研究大都基于短语结构树，基于依存关系的越南语语义角色标注研究较少，首要原因在于高质量的越南语依存树库资源的缺乏。

Liao 等[162]利用观察到的句子语义增量特征直接对句子的语义函数进行建模。为了能够利用现有的神经网络语言模型来近似语义函数，首先在语义函数上实现一阶泰勒展开，得到一个差分语义模型，然后作为子任务添加到 BERT 中，进行自我监督的微调。

Tuyen 等[163]提出了一种越南语语义信息检索模型，用于检索与查询相似的文本。在该系统中，语义分析是识别句子的语义依赖图，检索过程利用这些语义依赖图计算文档的相关性。为了识别句子的语义依赖图，利用越南语词典本体研究了转换规则在依赖分析中的应用。对于排序检索结果，将 Jaccard Tanimoto 距离应用于排序函数。

Dien Dinh et al.(2016)提出了一种对越南语句子释义进行分类的

新方法，该方法部署了预训练模型来利用语义上下文和语言知识，从而在识别过程中提供进一步的信息。

4. 机器翻译

与英文相比，汉语、越南语对句法结构以及语法规则的限定性远不如英文，这就给汉、越自然语言处理带来了一定的困难，同样直接影响到汉越机器翻译任务。

徐毓等[164]提出了一种基于深度可分离卷积的汉越神经机器翻译方法，根据越南语的语言特点,将越南语切分为词、音节、字符、子词 4 种不同的粒度并利用深度可分离卷积改进神经机器翻译模型，它通过增加深度可分离卷积神经网络对模型输入的不同粒度序列进行卷积运算，所以相比传统卷积降低了模型的理论计算量。经过实验，该方法在越南语 4 种不同翻译粒度上均取得最佳效果，一定程度上提升了汉越神经机器翻译性能。王振晗等[165]提出了一种融合源语言句法解析树的汉越神经机器翻译方法，利用深度优先遍历得到源语言的句法解析树的向量化表示，然后将句法向量与源语言词嵌入相加作为输入训练翻译模型。在汉-越语言对上进行了实验，其结果相较于基准系统，获得了 0.6 个 BLEU 值的提高。在此基础上，普浏清等[166]提出了一种基于依存图网络的汉越神经机器翻译方法，利用依存句法关系构建依存图网络并融入神经机器翻译模型中，在 Transformer 模型框架下引入一个图编码器对源语言的依存结构图进行向量化编码，并利用多头注意力机制将向量化的依存图结构编码融入到序列编码中，而在解码时利用该结构编码和序列编码一起指导模型解码生成译文。实验证明，在汉越翻译任务中融入依存句法图可以提升翻译模型的性能。

Tran et al.(2016)提出了一种在基于短语的英语到越南语统计机器翻译方法，它基于依存分析器自动学习重新排序规则作为预处理步骤的方法。通过依存分析和从训练特征丰富的判别分类器中提取规则

来重新排序源端句子。经过实验，该方法表现优于基于统计的机器翻译系统。Jiang 等[167]实现了一个在英越数据集上的原创 seq2seq 模型实验，通过加入注意力机制并对比模型的结果，结果表明注意力机制在越南机器翻译任务上表现出良好的性能。

5. 拼写检查

越南语中的字母和声调符号来自于拉丁文字和印欧语言。但是越南语又不像印欧语言那样有多余不发音的字母存在，它是由一个一个的字母构成，每个字母都有自己的读音。因此，如果把声调读错了，那么它相应的词语也会写错。越南语本身拼写规则的复杂性，这主要是由于各地的发音方法不同，而且有时候区别非常之大，拼写错误现象一直存在。

越南语拼写检查方法主要是基于越南语词汇语料库。My, Le Hoang Thi 等[169][168]提出了基于 Ede 音节模型的拼写检查 Ede 音节。目的是基于 Ede 音节模型构建 Ede 音节拼写检查算法。该算法用于将不在 Ede 词汇表中的单词分为两个结构组：正确和不正确的音节。Nguyen 等[169]将语言模型与字典和越南语词汇结构相结合，通过检测非标准单词以及拼写错误并纠正它们来规范越南语推文。Nguyen et al.(2016)提出一种利用 NER 学习模型进行规范化推文方法，归一化步骤检测推文中的拼写错误，并使用改进的 Dice 系数或 n-gram 进行纠正。

6. 情感分析

越南语的情感分类是越南语事件观点分析的基础。由于越南语资源匮乏，标注困难，越南语语言情感分类研究进展缓慢。

Thanh et al.(2017)提出了一种混合模型，该模型基于层次狄利克雷过程 (HDP)和结合了情感词典的支持向量机 (SVM) 方法进行组合。实验表明，该模型的平均准确率接近 87%。Quan-Hoang et al.(2017)将 CNN 和 LSTM 结合来处理越南语情感分析的任务，在 VLSP 语料

库上评估，该模型优于 SVM、LSTM 和 CNN。

近年来，BERT 预训练模型也逐渐被应用于越南语的情感分析任务中[170]。 Phan 等[171]创建了越南智能手机反馈数据集 UIT-ViSFD，由 11,122 条用于移动电子商务的人工注释评论组成。同时提出了一种基于 Bi-LSTM 架构和 fastText 词嵌入的方法，用于越南语方面的情感任务。实验表明，该方法在 aspect 任务中实现了 84.48% 的最佳性能(F1 分数)。

7. 自动摘要

作为低资源语言，越南语文本自动摘要的研究成果不多，可用的公开数据集较少。

Tu-Anh et al.(2012)将文本文档建模为加权无向图来生成与共同主题相关的多个越南语文档的提取摘要。首先构建无向图，然后通过 PageRank 算法，计算句子的显著分数并排序，根据最大边际相关性进行选择形成摘要。实验证明，所提出的技术在参考系统上有效。Thu H et al.(2013)提出基于摘录式的越南语文本摘要方法，通过神经网络学习结合降维特征来克服构建术语集时的成本并降低计算复杂度。实验表明，该方法在降低计算复杂度方面确实有效，并且优于之前提出的一些方法。Nguyen et al.(2016)提出一种使用网格模型和动态规划来计算 n-gram 以生成最佳句子压缩的新方法，由压缩后的句子组成文本摘要.实验表明，该方法非常有效，所生成的摘要文本在语法上、连贯性、简洁性方面较好。

在自动摘要数据集方面，Nguyen 等[172]发布了第一个大型的越南语文本摘要数据集。Tran 等[173]构建了一个用于多文档摘要的大型高质量越南数据集 ViMs，并通过各种指标验证了数据集的可靠性，结果表明 ViMs 数据集适用于训练和评估多文档摘要系统。

8. 命名实体识别

越南语命名实体识别是很困难的一项任务，其原因包括：1)实体

复杂。越南国家受多种文化的影响，在实体命名方面显示出命名实体的多样性和复杂性；越南地名命名广泛，主要分为基本地名和复合地名；越南语实体拼写多样化，比如：胡志明(tphcm,hồ chí minh,hochimin.)等；地名中同时含有数字出现，比如 3 号国道(“quốc lộ số 3”)；同时越南语和其他语言一样都存在外来词现象等；2)越南语有其独特的语言特点。越南语是孤立语，没有丰富的形态变化；越南语词是由一个或多个词素构成；越南人名和中国人名类似，唯一不同在于人名存在垫字，例如“Nguyễn Thị Tuyết”阮氏雪，常见的垫字有“文”(Văn)、“妙”(Diệu)、“女”(Nữ)、“玉”(ngọc)、“氏”(Thị)等；越南地名各音节首字母大写；比如：云南(Vân Nam)；非汉越外国地名，首字母大写，音节内部使用“-”连接，比如：Oen-linh-tợn；越南语机构、团体名称一般第一个音节首字母大写(词组除外)等。以上问题给越南语命名实体识别带来极大的困难与挑战。

闫丹辉等(2014)分析了越南语命名实体的语言学特点，在对其进行分类和形式化表达的基础上，提出了一种基于规则的越南语命名实体识别方法。结果表明，该方法准确率可以达到 90%以上，但是召回率不足 80%，其原因在于难以手工总结出所有可能的规则，制约了系统召回率的提高。刘艳超等(2016)结合实体库提出了越南语命名实体识别混合方法，首先根据越南语的语言和实体特点，选取有效的局部特征和全局特征，应用最大熵模型进行越南语命名实体识别；然后根据其制定的命名实体的规则进行越南语命名实体识别，再结合两者的识别结果，以规则为主，统计为辅原则，最后经过人工校对，把获取到的正确标记的实体加入到实体库，动态扩增实体库,为规则制定和特征选取提供丰富的语料和依据。经过实验该方法能够有效地结合规则与统计的方法优点，互相弥补不足。

基于支持向量机 (SVM) 的越南语命名实体识别模型(Tran et al.，2007)是较早的一种机器学习方法在越南语命名实体识别中的应用。

作者分析了越南语的构词特征及词形特征，在此基础上提出了一个基于 SVM 的越南语命名实体识别模型。实验结果表明，该系统对越南语人名、机构名、地名识别的准确率分别达到了 92.91%、85.16%、89.13%，召回率分别达到了 87.09%、77.11%、88.75%。

Pham et al.(2017)将双向长短期记忆 (Bi-LSTM)、卷积神经网络(CNN)、条件随机场 (CRF) 的组合模型应用于越南语命名实体识别任务，在标准数据集 VLSP 上 F1 值得分为 88.59%，达到了 VLSP 数据集 NER 比赛的第一梯队表现。BI-LSTM-CRF 组合模型并结合 POS 特征[174]，在 VLSP 数据集的 NER 任务中 F1 值达到 91.19%。

9. 未来挑战和发展趋势

越南语属低资源语言，兼有中西两种语言的特点，自然语言处理更为复杂，受科研工作者关注度较低，高水平研究成果较少，越南语自然语言处理领域仍存在诸多问题。

词汇层面，存在突变类型、兼词现象等，有众多一词多义以及同形异性的情况，由词汇层面的歧义所造成的机器在做分词和词性标注上的困难是很难克服的，仅依靠建立规则库来解决歧义问题必然会造成规则库规模过大、执行效率低等问题。

语法层面，越南语没有特定分隔或标志性词语，存在着结构方面的歧义。在机器分析这类结构时可用多种句法树来表示，通过经验主义的方法，利用大规模语料库构建语言模型对每个词之间的转换概率进行计算，从而选择一条概率最大的路径，进而寻找出最符合人们思维认知以及上下文语境的结构。

语义层面，民族之间风俗习惯的差异性必然会导致两种语言在表达上的不同。需要结合语境来考察，而语境信息如何融入系统中则需要知识库或者统计概率的支持。但目前的系统都无法做到对语境的充分理解与应用。

机器翻译层面，越南语存在着大量的长距离依赖现象。现阶段主

流的机器翻译系统大都基于长短时记忆网络，但长短时记忆网络对于长距离依赖问题的解决是有限的，其对过长的依赖问题也不能很好的解决。

越南语自然处理未来发展途径是建设一定规模的高质量语言资源，并融合越南语语言特征改进语言计算模型。

首先建设高质量语料库、字典以及各类规范标准数据集。尽管一些团队建立了一些越南语语言资源，但普通存在质量较低、不规范现象，缺少公开的标准数据集，难以有效支撑词法分析、句法分析、翻译等自然语言处理算法。人工标注语料存在工作量大、效率低等问题，大规模语料仅靠人工标注存在很大困难，如何利用少量人工标注语料，借助人工智能、机器学习方法构建大规模标注语料对于推动越南语自然语言处理具有重要意义。

其次越南语自然语言处理各领域普遍采用的方法是借助通用语言计算方法直接迁移到越南语语言信息处理中，尽管取得一定效果，但与通用语种的效果相比以及实际应用需求存在很大差距。针对越南语特点，提取语言特征，并将语言特征融入现有的性能较好的语言计算模型中，语言知识融合进去将是进一步突破的关键所在，融合越南语语言特征的分析及应用是越南语自然语言处理发展的必然趋势。

3.2.4 泰语

1. 语料库

目前，泰国主要的泰语语料库包括泰国国家电子计算机技术中心和朱拉隆功大学语言学系共同开发的泰国国家语料库(TNC)、HSE 泰语语料库、以及 SEAlang 泰语图书馆。其中泰国国家语料库(TNC)目标语料容量为 8000 万泰语词汇，截止到 2022 年 1 月，已收录语料 3300 万泰语词汇。在词语类型的划分处理规则上遵循 TEI(Text Encoding Initiatives)和 CES(Corpus Encoding Standard)，语料来源类型和比例参照英国国家语料库 BNC(British National Corpus)，选取原则

为语料的内容、建立时间和媒介类型，以保证所选取语料的多样性原则。HSE 泰语语料库收录泰语语料 5000 万词汇，SEAlang 泰语图书馆资源包括单语和双语词典、单语和双语语料库。以上三个语料库均对公众提供检索服务，语料来源主要类型包括报纸、杂志、文学作品、学术文献、法律文献、演讲文稿等，主要用于语言教学和研究。

随着机器翻译的迅猛发展，双语平行语料库的构建也日益重要。当前国内外泰汉平行语料库构建研究主要集中在国内。孙帅强[175]以互联网双语新闻网页为挖掘资源，进行了汉泰双语可比语料库相关技术方面的研究工作，具体包括汉泰双语可比语料库的构建、基于汉泰可比语料库的平行句对抽取、基于汉泰可比语料库的命名实体互译对抽取。此外，还有学者在进行汉语—泰语双语语料挖掘研究，此类研究有利于为汉泰双语语料库的建设夯实基础。张金鹏[176]围绕基于跨语言语料的汉泰词分布表示，汉泰双语实体对齐方法和汉泰双语新闻话题发现三个问题展开了相关研究，旨在自动挖掘汉泰双语新闻话题。侯中熙(2016)提出融合新闻要素的新闻文本相似度计算方法，研究汉—泰双语新闻文本相似度计算。

整体而言，目前泰语单语语料库的建设工作主要由泰国朱拉隆功大学自然语言处理实验室等相关泰国机构在完成，中国国内针对汉泰双语平行语料库也在积极开展前期研究工作，国际上也出现了一些泰语相关的平行语料库，但整体规模以及质量仍有较大的提升空间。

2. 词法分析

(1)分词

传统的泰语分词存在一定的局限性，如特征定义复杂等。近年来，由于深度学习强大的特征学习能力，深度学习的方法也被应用于泰语分词。目前研究者多采用多层神经网络，相比传统的机器学习方法，取得了更好的分词效果，而且分词速度也有较大的提升。Lapjaturapit T[177]等提出的泰语分词结合了泰语字符簇级(TCCLevel)和字符级信

息。该方法基于 TCC 和基于特征的 BiLSTM 神经网络，以便跟踪有用的信息。然后，使用模型的输出概率来表示输出候选的数量。最后，通过实验证明无论是基于 TCC 的方法还是基于字符的方法，该模型都达到了最好的性能。鉴于现有的深度学习系统速度较慢，Chormai P 等[178]提出了一种快速、准确的神经网络泰语分词方法 AttaCut，它使用扩展的 CNN 过滤器来捕捉每个字符的环境，并使用音节嵌入作为特征。系统运行速度相比之前至少提高了 5.6 倍，并在某些领域上超过了以前最先进的系统。Seeha S 等[179]使用来自双向语言模型的无监督预训练字符表示来改进有监督的分词系统 ThaiLMCut，证明了提出的半监督方法在没有任何复杂微调方法的情况下，总体上可以提高分词性能，特别是在标注数据量有限的情况下。另外实验结果表明，F1 得分最高增加了 2.02%，并且在标准基准 InterBEST2009 上获得 98.78% 的 F1 分数。

国内也有学者进行泰语分词的相关研究。陶广奉等[180]提出了基于上下文字符信息的泰语神经网络分词模型，该模型借助词分布表示方法，训练泰语字符表示向量，利用多层神经网络分类器实现泰语分词。在 InterBEST 2009 泰语分词评测语料上的实验结果表明，所提方法相较于其他基线模型取得了更好的分词效果，分词准确率、召回率和 F 值分别达到了 97.27%、99.26%及 98.26%。吴辉文[181]提出并实现了一种基于序列到序列的泰语分词模型：Glove-Seq2Seq。该模型通过一个双向长短期记忆(Long Short-Term Memory，LSTM)网络和一个单向门控循环单元(Gate Recurrent Unit，GRU)网络，将一个输入序列转换成另一个输出序列。Glove-Seq2Seq 与当前常用的多个分词模型在 4 个不同领域的数据集上进行了对比验证，结果表明，提出的分词模型简单有效，具有较强的领域适用性，在数据资源有限的情况下，仍可以达到理想的效果。

(2)词性标注

泰国学者 Sornlertlamvanich V 等[182]介绍了构建 ORCHID 泰语词性(POS)标记语料库的过程。ORCHID 是第一个构建 Thai POS 标记语料库的项目。它不仅限于泰语和 POS 标记语料库。语料库标记为三个级别：段落、句子和单词。由于泰语文本中没有明确的单词/句子边界、标点符号和变形，必须在标记 POS 之前将段落分成句子。该研究应用概率三元模型同时进行分词和词性标注。音节构造规则还用于减少计算概率的候选数。POS 分配中的问题被形式化以减少在相似 POS 情况下发生的歧义。国内，赵世瑜[183]提出使用隐马尔科夫模型和条件随机场模型实现了泰语词性标注。并提出了结合词向量的神经网络泰语词性标注方法。实验结果表明，融入词向量的神经网络泰语词性标注的准确率优于隐马尔科夫和条件随机场词性标注方法。目前，开源的泰语词性标注工具包有 PyThaiNLP 以及 TLTK。

3. 句法分析

目前，汉语、英语等语言采用基于传统的依存句法分析的方法研究句法解析相对成熟，但是传统的句法分析方法依赖于大规模标注语料库和制定复杂的特征模板，人工标注语料库和制定特征模板费时费力，泰语作为稀缺资源在研究上很难正常开展。但是汉语和泰语同属汉藏语系，两种语言在句法上有很大的相似性。因此，在依存句法解析的研究上泰语也可以借鉴汉语的研究。

陶广奉等[184]提出基于跨语言迁移学习的方法研究缺乏语料资源的依存句法解析，并完成了如下工作：(1)基于汉泰平行句对语料的神经网络双语词分布表示方法，实验结果表明词分布表示的准确率达到 82.60%。(2)基于迁移学习的泰语依存句法解析方法。在双语词分布表示方法的基础上，运用 40,000 句汉泰平行句对语料，通过从汉语中迁移特征的方法对泰语依存句法分析进行研究。研究所提出的神经网络泰语依存句法解析模型，在依存弧准确率、标识准确率和句子根节点的准确率分别达到 79.28%、75.01%和 91.25%。(3)泰语依存句

法分析系统的可视化。当前系统主要采用 Java 语言进行开发，输出 CoNLL 格式的依存语句，同时借助 DependencyViewer 工具进行界面化显示。

泰语句子相似度计算也是信息处理领域的问题之一。泰语句子相似度计算在多个领域有广泛的应用，特别是泰语机器翻译以及专家系统领域，相似度计算是不可或缺的部分。关于汉-泰跨语言句子相似度计算方法的研究，洪玄贵[185]从泰语句子的关键词、知网的语义、WordNet 语料库的语义三个方面的特征进行了研究。冯银汉(2019)从三个方面对汉-泰跨语言句子相似度计算方法展开相关研究：(1)在泰语单语言的句子相似度计算方面，提出基于词性和词向量的泰语句子相似度计算方法。(2)提出基于不对等语料的汉-泰跨语言词语的相似度计算方法。(3)提出基于句子嵌入的汉-泰跨语言句子相似度计算方法。

4. 机器翻译

在机器翻译方面，Lalita Chinese-Thai Machine Translation 中泰双语翻译模型，翻译效率约为谷歌翻译的 60-70%，在所有翻译方向均优于 NECTEC 的 AIforThai (xiaofan)。此外，由泰国人工智能研究院开发的机器翻译模型——English-Thai Machine Translation Models，主要被训练用于从英泰句子对数据集中翻译 2 种语言对，即泰语→英语和英语→泰语(scb-mt-en-th-2020)，其中包含超过 100 万个句对。从测试结果中发现，翻译模型 Thai→English 和 English→Thai 可以执行相当于或优于 Google Translation API 翻译系统的翻译(截至 2020 年 5 月测试)。此外，还有 PyThaiNLP 以及 thai2nmt 也集成了英泰机器翻译模型。

5. 自动文本摘要

Thattinaphanich S 等[186]采用预训练语言模型 BERT 进行抽取式摘要，Liu Y[187]扩展了一个最新的抽取摘要模型，提出了

AREDSUM-SEQ 和 AREDSUM-CTX，AREDSUM-SEQ 通过序列生成模型对句子进行联合评分和选择，AREDSUM-CTX 通过单独的模型学习平衡显著和冗余。实验结果表明，AREDSUM-CTX 的性能优于 AREDSUM-SEQ 和所有其他强基线，这表明冗余度的减少有助于提高摘要质量，并且在句子评分过程中，显式建模冗余度的效果要好于联合使用显著度。

Seq2Seq 模型在文本摘要方面取得了巨大成就。然而，Seq2Seq 模型往往需要大规模的训练数据才能达到有效的结果。泰文摘要的进展还远远落后，大规模数据集的缺乏使泰语文本摘要处于起步阶段。

用于泰语文本摘要的大规模数据集主要是 ThaiSum 以及 TR-TPBS。ThaiSum 是一个用于泰语文本摘要的大型语料库，这也是最大的泰语文本摘要语料库，该语料库从几个在线新闻网站获得，即 Thairath、ThaiPBS、Prachathai 和 The Standard。该数据集由记者撰写的超过 350,000 条文章和摘要对组成。此外，TR-TPBS 是一个中等大小的数据集，是一个多用途的 NLP 基准测试，特别是对于泰语。此数据集是从 Thairath(TR)和 ThaiPBS(TPBS)新闻网站抓取的。此语料库的主要目的是用于泰文文本摘要。

泰语的文本摘要大都是基于上述两个泰语文本摘要语料库进行训练，除此之外，还可以基于语料库训练的多语言模型 Multilingual T5(mT5)，该模型是一个大规模的多语言预先训练的文本到文本转换器模型，按照与 T5 类似的方式进行训练。

6. 命名实体识别

命名实体识别是从给定的文本中提取命名实体的任务。它标识每个实体的跨度，并将标识的跨度分类到实体类别中。Thattinaphanich S[188]提出了一种基于词和字符表示的 BiLSTM-CRF。首先，通过将一个句子标记化为一串单词来准备文本。然后，进行单词表示和 BiLSTM 字符表示。最后，利用递归神经网络和 CRF 相结合的方法

来学习文本序列，并提取知识来建立 NER 识别来克服这一问题。该模型通过 Facebook 小组 ThaiNLP 的开源语料库进行了评估。模型的准确率、召回率和 F1 分别为 91.79%、91.51%和 91.65%。

由于泰语构词方法和语法规则复杂，针对这一问题，昆明理工大学王红斌等人[189]将泰语命名实体识别任务转化为对泰语句子中的词汇序列进行标记。结合泰语的语言特点，选择合适的泰语上下文特征，分别使用隐马尔科夫模型和条件随机场模型在泰语实体识别训练语料上进行了模型构建，并在测试语料上对所构建的序列标注模型进行了实验验证。实验结果表明，使用隐马尔科夫模型和条件随机场模型进行泰语人名、地名、机构名实体识别是可行的,并取得了较好的效果。吴辉文[190]针对泰语文本中的人名、地名和组织机构名等实体抽取任务，将其展开为实体识别和实体分类两个任务进行研究。首先构建了一种基于标签注意力网络进行逐层改进的模型，该模型分为两层结构，每一层由对序列信息进编码的 BiGRU 编码子层和对标签信息进行推理的标签注意力推断层组成。该模型与多个常用的实体识别模型进行对比，最终的实验结果验证了模型的优越性能，更适用于泰语实体识别任务。针对泰语实体分类任务，构建了基于注意力增强的 Bi LSTM 神经网络分类模型，该模型通过将 Bi LSTM 神经网络结合注意力机制，有效地增强了实体分类效果。

在泰语命名实体识别语料库上，Buaphet W[191]创建了泰语首个嵌套命名实体识别(N-NER)数据集，该数据集使用覆盖广泛用例的广泛标记集进行标注，该数据集也是最大的非英语 N-NER 数据集，也是第一个具有细粒度类别的非英语数据集。该语料库解决了泰国自然语言处理的数据稀缺问题。

开源的 PyThaiNLP、Thai-NNER 以及 TLTK 泰语工具包也能实现泰语命名实体识别功能。

7. 语言模型

将泰语集成到训练数据集中的多语言语言模型(例如 mBERT 和 XLMR)存在一些局限性，例如：模型训练同时对 100 多种语言的数据进行学习，从而无法具体到语言使用模式，或添加出现在特定泰语数据集中的各种主题，例如，XLMR 依赖于单个网站爬网， mBERT 仅依赖于维基百科。对此，Lowphansirikul L[192]提出了 ThaiFit 模型，以前称为 thai2vec 模型。ThaiFit 模型是一个 ULMFit 模型，该模型作为开源的 pyThaiNLP 库的一部分，在泰语维基百科上针对泰语文本分类进行了训练。在文本分类上，thai2fit 击败了 Facebook 的 fast Text 和谷歌的 Bert，是当前泰语文本分类的最先进技术。此外，BERT-th 也是当前主流的泰语预训练模型。BERT 是一个预先训练的无监督自然语言处理模型，它为微调做好了准备，以便显著地执行 NLP 下游任务。为了在泰国计算语言资源极少的情况下提供研究机会，BERT-th 提出了基于 BERT-Base 结构的仅限泰语的预训练模型。在此基础上，Bi K[193]等提出了 WangchanBERTa 的训练模型并指出，对于资源相对较少的语言(如泰语)，模型的选择仅限于基于更小的数据集训练基于 BERT 的模型或微调多语言模型，这两者都会产生次优的下游性能。此外，大规模的多语种预训没有考虑到泰语的语言特点。为了克服这些限制，WangchanBERTa 在一个大型的、重复数据消除的、干净的训练集(总大小为 78 GB)上预训练了一个基于 Roberta-Base 架构的语言模型，该训练集来自社交媒体帖子、新闻文章和其他公开可用的不同领域的数据集。总而言之，在大型的数据集(如泰国的文本分类数据集)上训练可以产生更好的下游性能。

8. 未来挑战和发展趋势

目前，泰国的主要研究团队集中在朱拉隆功大学以及泰国电子与计算机技术中心(NECTEC)，此外，泰国农业大学亦有相关研究。2019 年 5 月 17 日，朱拉隆功大学自然语言处理实验室发布了泰国文学语

料库；2019 年 12 月，泰国电子与计算机技术中心(NECTEC)与泰国开泰银行、泰国朱拉隆功大学宣布共同合作开发应用于金融和商业领域的自然语言处理技术。

在国内，泰语自然语言处理研究多见于昆明理工大学。整体呈现研究覆盖范围广，以基础研究为主的特点。基础研究主要集中在语料库构建、分词、词性标注、句法分析、句子相似度计算、命名实体识别等，应用层面主要关注机器翻译、自动文本摘要等。2021 年 11 月，云南省人工智能重点实验室发布了云岭机器翻译平台，该平台根据一系列神经机器翻译方法，实现了中文与越南语、缅甸语、泰语、老挝语、英语、日语等 108 个语种的双向神经机器翻译。

综上所述，目前泰语自然语言处理技术已经初露雏形。在应用层面上，泰国正在加强自然语言处理技术与金融领域等其它相关产业的结合，国内也在积极开展针对泰语自然语言处理应用的相关研究，包括：机器翻译、情感分析、自动文本摘要以及文本生成技术等。但无论是底层技术研究亦或是具体应用，都需要依赖高质量、大规模的标注语料，然而，目前泰语语料库规模与其它通用语种相比仍有较大差距。因此，在下一阶段，加强泰语的通用语料库以及专用语料库构建，探索针

对低资源语言的计算模型及语言处理技术将是重点研究方向。

3.2.5 老挝语

1. 语料库与机器翻译

目前，老挝语语料库还十分欠缺，老挝语-汉语的研究团队在很大程度上依赖团队构建的未公开的小规模语料，例如使用实验室历年积累的小型人工标注语料库[194]、汉语与老挝语双语语料库[195]、东南亚语言信息处理平台[196]等进行自然语言处理研究。

Tien 等[197]提出一种改进的句对齐方法。该方法使用具有句子长度比的嵌入模型来对齐双语文档的句子，减少需要考虑的候选句对齐

语言对，在越南语-老挝语句对实验上的精度达到 95.74%。

针对老挝语等低资源语言在测试题上缺乏大规模的高质量数据和知识来源问题，Qiu 等[198]提出了一个神经模型框架和一套干扰项生成策略，采用 3 种策略生成 5 种针对完型填空的干扰选项，为未来的研究提供了一个老挝语完型填空测试语料库。

由于受到中老语言对平行语料库规模的严重限制，中老神经机器翻译(NMT)任务表现不佳。Yu 等[199]发现泰国-老挝语在跨语言方面有很大的相似性。根据这些特点，提出了一种新的 NMT 方法。首先训练中泰和泰老 NMT 模型，其中泰语被视为中枢语言。然后采用迁移学习策略分别从两个训练模型中提取编码器和解码器。最后，将编码器和解码器组合成一个新的模型，然后基于一个小型的汉语-老挝语平行语料库进行微调。与 transformer 基线模型相比，该方法提升了 3.62 BLEU 值。

2. 词法分析

作为低资源语言的老挝语，国内外老挝语词性标注方面的研究仍比较薄弱。杨蓓等[200]提出了基于半监督学习的老挝语词性标注方法，结合整数规划和二阶隐马尔可夫模型，利用少量标注词典和未标注语料资源实现高质量的老挝语词性标注，其准确率达到 89.8%。王兴金等[201]则在半监督隐马尔科夫模型的基础上融合了词预测模型，以解决未登录词词性标注问题，并采用规则和统计相结合的方法来提高 HMM 标注精度与速度。此外，王兴金等[202]还通过分析老挝词结构，构建了结合词性标注损失和主辅音辅助损失的多任务老挝语词性标注模型，在有限标注语料下获得了更好的表现(93.24%)。

彭骁男等[203]根据老挝语人名地名语言学中句法与词法的相关特征，使用 Bi-LSTM 进行词语字符级向量训练，将字符级特征向量和词向量组合拼接成组合向量。然后将老挝语地名做状语后置的句法特征，通过构建特征向量输入到 CRF 中进行命名实体识别训练。结果

表明，融合了多特征的老挝语命名实体识别模型的准确率、F 值得到4%左右的提高。

3. 句法分析

老挝语依存分析研究包括构建老挝语依存标注体系和老挝语依存树库，为老挝语的机器翻译和句法解析等研究提供支持。

殷若尘[204]提出了一种借助汉-老双语词对齐语料构建老挝语依存树库的方法。在已经获取汉-老双语词对齐平行语料的基础上，首先对平行语料中的汉语句子进行依存句法分析，然后结合老挝语自身语言特点，在依存句法规则的基础上将汉语句子的依存关系通过汉-老双语词对齐关系映射到老挝语句子中，最终生成老挝语句子的依存树。李炫达[205]提出一种融合句子结构特征的汉老双语句子相似度计算方法。汉老双语句子具有相似的句子结构特征，通过构建特征模板获取汉语和老挝语对应的句子结构特征，获取更多的语义信息，将双语词向量映射到共享语义空间以减小语言间差异性，最终构建汉老双语句子相似度计算模型。

4. 篇章分析

李思卓等[210]提出一种基于互译特征词对匹配，构建老-汉双语句子相似度计算方法，改进了传统的依赖于词形词序通过计算相同词个数和共有单词的位置信息的相似度计算方法，充分考虑了老挝语和汉语句子中的词汇互译信息、相似概率，避免了由于特征词位置导致的精度丢失。

Chen 等[206]针对老挝语文本分类研究很少的情况，提出了一种基于 KNN 的老挝新闻文本分类方法。首先对老挝新闻文本进行预处理和特征提取，然后通过 KNN 分类器调整参数，最后在数据归一化和数据降维中进行处理，从而提高分类效果。

何阳宇等[207]提出一种基于双向长短期记忆网络和多头自注意力机制的军事领域实体关系抽取方法。针对老挝语语料匮乏问题,提出

了“硬匹配”和“软匹配”的思想，在完成语料获取和预处理的基础上，利用预定义的关系词表进行“硬匹配”，之后再通过词典匹配和相似度计算相结合的方法进行"软匹配"，以提高关系类型的泛化能力，进而构建了关系抽取标注语料库。

针对中国-老挝短文本相似度计算任务，Li 等[208]提出结合中老双语者共同的词性特征，从有限的训练语料库中获得更多的语义信息。同时，作者还通过 Poly 网络获取文本匹配信息，进一步提高语义向量表示的准确性，构建了中老两语双语短文本的相似性计算模型。

郭雷等[209]提出一种融合词语多特征的汉老短文本相似度计算方法，首先利用双向长短期记忆网络(BiLSTM)和卷积神经网络(CNN)分别提取汉老词语的形态学特征，将词向量拼接上形态学特征向量、词性向量、词性权重向量，然后利用 BiLSTM 和 CNN 提取汉老短文本的上下文特征和局部语义特征，加入交互注意力机制，最后计算汉老特征语义向量的相对差和相对积，将其结果拼接并输入到全连接层得到汉老双语短文本的相似度分数。

谭琪辉等[211]提出了一种融合文本特征的汉老双语句子相似度计算方法，并构建了句子相似度模型。该模型将汉语、老挝语的词性、数字共现等文本特征与 GloVe 预训练词向量融合，以此丰富句子特征，提升模型计算准确率。其次，由基于自注意力的双向长短时记忆网络组成多层孪生网络来提取长距离上下文特征和深层次语义信息。最后,采用迁移学习的方法将通用模型参数初始化，并使用不同的微调参策略增强模型的泛化能力。

受跨语言分布表示学习的启发，何力[212]使用汉老双语对齐句向量预训练深度典型关联分析 DeepCCA(Deep Canonical Correlation Analysis)模型联系双语句子并计算其相似度。首先分别向量化表示双语句子，然后使用预训练的 DeepCCA 模型，将双语句向量映射到新的空间内，最后在新空间内利用映射后句向量的余弦距离来计算汉老

双语句子相似度。实验证明本方法能有效计算汉老双语句子相似度。

杨志婥琪等[213]提出了一种融合字符形状特征的多任务老挝语文字识别后纠错方法。该方法引入基于长短期记忆网络的 seq2seq 模型架构，将老挝字形特征融入模型以辅助模型对相似字符替换错误的纠正；使用语言模型对解码端预测的文本序列与原始文本进行重排名，得到最佳候选；同时，采用多任务学习的方式，以错误检测辅任务优化模型纠错效果，此外，该文以数据增强的方式扩充数据集。实验结果表明，该方法使老挝文字识别的字符错误率低至 7.94%。

5. 存在的问题和展望

老挝语目前非常缺乏相关的语料资源，而通过领域专家以人工方式难以在短时间内提供出有助于解决“一带一路”多语言信息需求的资源库或技术服务，目前的研究对于新闻资讯、经贸往来、跨境电商、旅游观光等领域的语言学习和应用的辅助平台具有很重要的现实意义。目前的老挝语自然语言处理研究在语料库构建和词法分析、文本相似度计算方面取得了一定的成果，而在句法分析、依存树库构建方面还有进一步研究的空间。

3.2.6 柬埔寨语

1. 语料库

柬埔寨语即高棉语。柬汉双语网站较少，所以通过网络获取一定规模且高质量的平行语料比较困难，本节介绍汉柬双语可比语料库构建方法的进展。

高棉语在词典获取方面取得一定的进展，Mangeot 等[214]开发了针对高棉语的多语言词汇系统。数据主要来自 Denis Richer 的法语-高棉双语词典从 Word 到 XML 格式的转换。生成的资源可通过 Jibiki 平台上的 REST API 在线获取，以进行查找、编辑、下载和远程编程。

在可比语料库构建方面，刘小慧[215]研究基于文本层次聚类的可比语料获取方法，引入了语义关系并对文本进行聚类分析，将文本建

模和文本聚类相结合，最后从聚类结果中获取可比语料，与基于词典的文本相似度计算方法进行聚类相比，效果更佳。

潘丽同[216]研究英柬双语平行句对的识别。为了从候选平行句对中识别出平行句对，构造了一个二分类的最大熵分类器。采用句子长度特征、词汇化比例特征、句子位置特征、符号特征等进行英柬双语句对分类器的训练。最后利用该分类器对英柬候选平行句对进行分类，从而确定出英柬双语平行句对资源。

李思远[217]研究基于双向循环神经网络的可比语料柬汉平行句对获取。与现有的通过机器翻译以及其他神经网络模型获取平行句对的方法相比，该方法提高了工作效率以及句对平行准确率，且不需要提供双语平行文本，解决了柬汉双语平行句对获取的问题。

Chi H 等[218]提出一种基于 Manhattan-BiGRU 模型的柬汉平行句对的获取方法。该方法将特征信息和双语词嵌入连接在一起共同用作输入。然后通过 BiGRU 网络将词嵌入编码为句子嵌入。最后根据曼哈顿距离算法计算句子嵌入的相似度，实现平行句对的获取。与其他基于神经网络模型获取平行句对的方法相比，实验结果表明该方法在可比语料库中可取得较好的效果。

2. 词法分析

针对高棉语，分词和词性标注方面的进展较大，而词形还原方面的究成果甚少。因此本节主要介绍高棉分词和词性标注的研究进展.

（1） 分词

潘华山等[223]针对高棉语分词及词性标注问题，提出一种基于层叠条件随机场模型的自动分词及词性标注方法。该方法由三层条件随机场模型构成：结合上下文信息与高棉语丰富的词缀信息构建特征模板，实现对高棉语句子中的词语进行自动标注词性。实验结果表明该方法能有效解决高棉语的分词和词性标注问题。

Tran Van Nam 等[222]提出了一种使用音节模型的划分成分簇的解

决方案，建立了一个高棉音节数据库。每个组件簇都由第一个字母和最后一个字母标记和定位，以标识整个音节。实验测试结果达到了高精度，消除了歧义，有助于解决分词问题并在高棉语处理中提高了效率。

Buoy 等[224]提出了一种使用单一深度学习模型的联合分词和词性标注方法，以便可以自发地执行分词和词性标注。使用公开可用的高棉 POS 数据集对所提出的模型进行了训练和测试。

（2） 词性标注

Sangvat 等[225]提出了一种使用条件随机场(CRF)进行高棉词性标注的替代方法。提取五组特征并将它们与 CRF 模型一起使用。提出的方法已经在 41,058 个单词和 27 个 POS 标签的语料库上进行了评估。

Sry S 等[226]描述了使用单个长短期记忆网络的实验研究。来自亚洲语言树库的数据集用于训练和测试模型。初步实验模型达到了 95% 的准确率。但是，需要进行更多的测试来评估模型并将其与不同的模型进行比较，以选择精度更高的模型。

3. 句法分析

当前，针对柬埔寨语的依存句法研究比较少，尚缺乏高质量、大规模的柬埔寨语依存树库和依存句法分析器的原型系统。

从已有的句法分析理论和方法出发，根据柬埔寨语的基本特征，借助于已经取得的研究成果提出在柬埔寨语研究更为贴合的句法依存分析法。徐璐[231]基于汉柬平行句对研究柬埔寨语依存句法分析，借助于词对齐将汉语分析得出的依存关系结果反映到柬语的句子中，确立约束规则对柬语依存关系做自我修复映射，进而得出数量众多的柬语依存标注语料。

在依存树库方面，Kann 等[232]构建高棉树库的半自动框架，从高棉语法书籍中的句子中提取高棉语法规则。在获得 Treebank 后生成语法规则概率。在实验中，注释树和提取的语法规则以定量和定性的

方式进行分析。根据三个验证的结果，Self-Consistency 的结果最好。

4. 机器翻译

Suraiya Jabin 等[219]开发了高棉语的在线混合机器翻译(MT)系统。实验结果表明，以英语为源语言，高棉语为目标语言的翻译是成功的。

Prasomsuk Sukchatri 等[220]提出了一种有效的泰语到高棉语机器翻译系统。通过考虑前一个单词，下一个单词和主语-动词一致性来应用单词重新排序，调整单词以获得可接受的输出。结果显示其效率高于 Google 和其他系统。

Marie B 等[221]介绍了高棉-英语翻译任务的有监督和无监督机器翻译系统。对于所有翻译方向，构建了有监督神经机器翻译系统(NMT)和统计机器翻译系统(SMT)。使用经过清理和规范化的单语数据，NMT 和 SMT 组合在四个翻译方向上表现最好。

5. 命名实体识别

高棉语命名实体识别是高棉语信息处理的一项基础工作。为了解决高棉语词法标注语料稀缺、高棉语命名实体缺乏明显标识特征的问题，黄淑慧[227]提出了一种融合高棉语实体特征的约束条件随机场模型的命名实体识别方法。通过设置的对比实验可知，这种约束的条件随机场模型在对高棉语进行命名实体识别时，效果比传统的条件随机场模型有了一定程度的提升。徐广义等[228]提出一种引入英柬跨语言特征的高棉语命名实体识别方法。实验结果表明，融入跨语言特征的条件随机场模型能有效地提升高棉语命名实体识别的效果。郭月江[229]提出融合跨语言特征的高棉语命名实体识别方法。利用较为成熟的英语命名实体识别技术，以英柬平行语料为桥梁，实现柬语的命名实体识别。针对 BiLSTM 神经网络模型输出没有考虑输出标签之间的顺序性，造成实体识别效果不良，谢俊[230]提出将 BiLSTM 神经网络模型的输出与柬埔寨的实体特征一起作为 CRF 模型的输入特征，利用 CRF 模型实现高棉语命名实体识别。实验结果表明该方法能够使

高棉语命名实体识别效果得到提高。

6. 情感分析

李小龙[233]在对高棉语文档级的基本算法进行优化改进的基础上，给出了三种特定情感分类模型。描述了情感分类系统的设计、情感分类系统分析的结果，介绍了性能指标，给出了实验模型的情感分类实验结果，对每个模型的性能给出了比较，讨论影响每个模型的准确性的因素。

Rifat 等[234]使用 FastText 和 BERT 语言模型来提取词嵌入，并进行了三种不同新闻文本情感分类的实验。实验结果表明，经过预训练和微调的 BERT 模型在高棉语中的情绪分析产生了出色的结果。

7. 未来挑战和发展趋势

针对高棉语的自然语言处理研究的一个基础难题是语言资源缺乏问题，因此后续需要进一步研究如何采用半自动或全自动的手段构建高质量、大规模的语料库。目前高棉语的词法分析研究(如分词、词性标注、命名实体识别)已有较丰富的研究成果，而句法分析、句法树构建和篇章层面的研究还有进一步推进的空间。

3.2.7 缅甸语

1. 语言资源构建

当前，国内外构建的缅甸语语言资源主要包括可比和平行语料及双语词典两类。

(1)可比和平行语料资源构建

张少宁(2019)提出了一种基于枢轴语言的汉-缅平行语料库构建方法。该方法以英语作为枢轴语言，通过构建汉-英-缅三者的公共语义空间，进而实现汉-缅平行句对抽取，最终实现汉-缅平行语料库的构建。毛存礼等[235]研究提出了基于 CNN-CorrNet 网络的汉缅平行句对抽取方法。该方法首先利用 BERT 得到汉语、缅语词向量表征，并将两种语言句子用卷积神经网络进行句子表征，以捕捉句子重要特征

信息。然后利用已有的汉缅平行句对作为约束条件，使用 CorrNet(相关神经网络)将汉缅的句子表征投影到公共语义空间，以此来保证两种语言跨语言表征的最大相关性。最后计算公共语义空间中汉语、缅语句子距离，并根据距离判断汉—缅双语句子是否为平行句子。李训宇等[236]研究提出了一种融合主题模型及双语词向量的汉缅双语可比文档获取方法。该方法的特点在于将跨语言文档相似度计算转化为跨语言主题相似度计算。首先，使用单语 LDA 主题模型分别抽取汉语、缅甸语的主题，得到对应的主题分布表示。然后，将抽取到的汉缅主题词进行表征得到单语的主题词向量，利用汉缅双语词典将汉语、缅甸语单语主题词向量映射到共享的语义空间，得到汉缅双语主题词向量。最后，通过计算汉语、缅甸语主题相似度获取汉缅双语可比文档。毛存礼等[237]提出了一种结构特征一致性约束的汉缅双语平行句对抽取方法。该方法是对基于孪生神经网络的双语平行句对抽取模型的扩展。首先，通过多语言 BERT 预训练语言模型在嵌入层将两种语言编码到同一语义空间，以此缩小语义空间中语言的差异。然后，分别对两种语言句子的长度特征进行编码，与孪生网络编码后的句子语义向量进行融合，增强平行句对在语义及结构特征上的表示，降低模型对语义相似但不平行句对的误判。

(2)双语词典构建

毛存礼等[238]提出了一种基于半监督的汉缅双语词典构建方法。该方法通过利用预训练语言模型来构建双语词汇的上下文特征向量，对基于可比语料和小规模种子词典的迭代自学习方法得到的汉缅双语词汇进行语义增强。李越等[239]提出了一种融合主题及上下文特征的汉缅双语词汇抽取方法。该方法首先利用 LDA 主题模型获取汉缅文档主题分布，并通过双语词向量表征将跨语言主题向量映射到共享的语义空间，然后抽取同一主题下相似度较高的词作为汉-缅双语候选词汇，再基于 BERT 获取候选双语词汇相关上下文的词汇语义表征

构建上下文向量，最后通过计算候选词的上下文向量的相似度对候选双语词汇进行加权得到质量更高的汉缅互译词汇。

2. 词法分析

当前，缅甸语词法分析领域的研究仍主要集中于分词、词性标注等基础性工作。

(1)分词

韩晓东(2016)对缅甸语的语言特点、缅甸语的分词模型构建以及缅甸语的编码转换等问题进行了研究分析，提出了一种基于规则的缅甸语音节切分方法。此外，还提出了一种融合音节特征的基于层叠条件随机场的缅甸语分词方法。林颂凯等[240]提出了一种基于卷积神经网络的缅甸语分词方法。该方法首先将缅甸语音节结构特征应用于缅甸语音节词向量特征分布式表示，然后基于卷积神经网络将音节及其上下文的特征进行融合，得到有效的特征表示，并通过深层网络的逐层特征优化自动学习到缅甸语分词的有效特征向量，最后利用 softmax 分类器来对构成缅甸语词汇的音节序列标记进行预测。马昌娥等[241]以 5000 个经过缅语专家人工分词的完整句子为数据集，实验对比了基于条件随机场(CRF)的缅语分词方法与基于正向最大匹配算法(FMM)的缅语分词方法，并用置信度、分词精度和分词速度评估分词性能。

(2)词性标注

李中伟(2017)提出了基于汉-缅双语可比语料的缅甸语词性标注语料库构建方法。该方法利用基于汉-缅双语词典及 WordNet 双语词语上下文向量相似度计算方法抽取汉-缅互译词，并利用双语词性映射的方法，实现缅甸语词性标注，构建缅甸语词性标注语料库。他同时提出了一种融合词典知识的缅甸语词性标注语料库构建方法。该方法在上一种方法获取的词性标注语料库中提取词语扩充英缅词典，利用英缅词典对缅甸语单语新闻分词文本进行词性粗标注，同时构建一

些规则库对未登录词和兼类词的词性标注的规则支持，并利用贝叶斯模型对兼类词进行词性消歧，最终完成缅甸语的词性标注工作，构建出缅甸语词性标注语料库。Kyaw Htet Minn 等[242]通过研究缅甸语的词形，实现了基于 n-gram 的分词，并提出缅甸语的语法词干规则和词性标注规则。Dim Lam Cing 等[243]提出了基于隐马尔可夫模型和形态学规则的缅甸语联合分词和词性标注。

3. 句法分析

早期，由于缺乏人工标注的句法分析数据，Ding 等[244]提出了一种以日语为枢轴的句法分析方法，原因是日语和缅甸语的句法结构相似。马文举(2019)利用英语的依存标注数据，通过迁移学习方法来研究缅甸语依存句法分析问题。他分析了缅甸语、英语在句法方面的差异性，提出了基于共享网络参数的缅甸语依存句法分析方法和基于迁移学习的缅甸语依存句法分析模型。

4. 命名实体识别

Aung Hla Moe(张家富)(2018)提出了基于汉—缅双语可比语料的双语实体抽取方法，并研发了汉—缅双语抽取原型系统。首先，抽取中文句子中的实体以及实体类别、位置、长度等特征，以此对缅文实体所在句子中的位置及长度进行约束。然后，基于缅语助词对缅语句子进行功能标记，并对缅语候选实体片段进行切分。最后，通过计算中文实体与候选缅语实体片段之间的相似度，选取相似度最大的候选片段作为对应的缅语实体。Aye Myat Mon 等(2020)构建了一个包含 8 万多个音译实例的缅甸语—英语命名实体词典，并使用基于统计和神经网络的方法评估了自动音译的性能。

5. 机器翻译

近年来，缅甸语机器翻译方面的研究已取得较大突破。Win Pa Pa 等[245]首次对 PBSMT、HPBSMT、树到字符串(T2S)、字符串到树(S2T)和 OSM 等五种应用于低资源语言的机器翻译方法做了比较研究，并

将其应用于双向翻译英语和(泰国、老挝、缅甸)之间有限数量的旅游领域数据。Ye Kyaw Thu 等[246]较早对缅甸语机器翻译进行了较大规模的研究，结果表明，基于短语的分层 SMT(HPBSMT)方法的翻译质量最高。Yi Mon Shwe Sin[247]提出了基于注意力机制的缅语—英语神经机器翻译系统，并构建了缅英平行语料库和缅语单语语料库。满志博等[248]提出了一种基于多语言联合训练的汉英缅神经机器翻译方法。该方法在 Transformer 框架下将丰富的汉英平行语料与较少的汉缅、英缅语料进行联合训练，训练过程中分别在编码端和解码端将汉英缅映射在同一语义空间，以此降低汉英缅语言结构差异性对共享词表的影响。同时，通过共享汉英语料训练参数来弥补汉缅、英缅语料缺失的问题。

6. 情感分析

Y M Aye 和 S S Aung[249]对缅甸餐厅及其食品的评论文本进行了研究，创建了情感数据集，旨在实现根据客户评论的情绪进行推荐。林颂凯(2018)通过双语词向量以及双语句子向量的表征，将汉语情感分析资源及方法应用在缅语中，完成缅语的情感分析。经过研究，提出了融合缅语音节特征的缅甸语词向量表示方法、汉缅双语句子级 embedding 语义表征方法和基于双语表示的缅甸语句子情感分类方法。Soe Yu Maw[250]以旅游领域的缅甸语评论文本为研究对象，提出了一种基于长短时记忆(LSTM)的情感分析方法。Hay Mar Su 等[251]对比研究了 Logistic 回归、SVM 和随机森林等三种机器学习(ML)技术在基于词向量表示法的 Facebook 缅文数据集情感分析中的性能差异。

7. 文本分类和自动摘要

目前，国内外在缅甸语文本分类和自动摘要等方面的研究仍处于起步阶段。文本分类方面，Myat Sapal Phyu 等[252]对卷积神经网络和长短时记忆(CNN-LSTM)联合应用于缅甸语文本分类的有效性做了研究和分析。文本摘要方面，Yamin Thu 等[253]提出了一种基于递归神

经网络的缅甸文章标题预测模型，并与序列对序列模型进行了比较，对该预测模型的性能做了评估。

8. 语言模型

王雍凯(2017)研究提出了一种融合上下文特征的汉—缅双语主题模型和一种融合语义扩展的汉-缅双语主题模型。其目的是通过主题模型获取可比语料，并构建语料库。其中，前者以双语 LDA 主题模型为基础，融合了文本的上下文特征。融合后的模型降低了高频词对文本主题分布的负面影响。后者以融合上下文特征的主题模型为基础，进一步融合了汉—缅语义扩展词典，通过对词典的解析和处理，构建了汉—缅语义的扩展集合。Aye Myat Mon 等[254]利用不同特征向量大小的词袋(CBOW)模型提取缅甸语新闻文档中的相似词。通过对词嵌入模型的分析，得到了高维向量比低维向量更好的基于相关性的单词聚类结果。

9. 语音合成与文语转换

除电子化文字语料资源建设、加工和分析研究外，近年来，国内学界有关缅甸语语音合成与文语转换的研究也取得了一定的成果。

语音合成方面，杨馨(2018)对缅甸语语音合成系统中的前端文本分析方法进行了研究。研究提出的音节边界规则、罗马化方案、分词方法及数字归一化方法，可基本满足开发缅语语音合成系统的要求。马昌娥(2019)通过构建发音语料库，研究并实现文本归一化、分词和文本注音。她根据 MLC(The Myanmar Language Commission)转写系统和 IPA(International Phonetic Alphabet)注音系统提出了基于声韵母拼接的文本自动注音方法。刘梦媛(2020)进一步研究改善了缅甸语文本分析方法和语音波形合成方法，提高了语音合成的自然度。经过持续的研究探索，刘梦媛、杨鉴[256]设计并实现了一个基于 HMM 的语音合成系统。首先，为使计算机按输入文本合成出正确的读音，提出并设计了缅甸语的注音方案，其中重点解决了缅甸语中的变音和变调

问题。然后，根据缅甸语的语音特点选取声母及带声调的韵母作为合成基元，并按此设计上下文属性和问题集。最后，基于 HTS 平台，完整实现了音子自动切分、HMM 声学模型训练及语音合成。实验结果表明，该缅甸语语音合成系统具有可行性，可作为后续研究的基线系统。

文语转换方面，杨馨、杨鉴[257]以开发缅甸语文语转换系统为目的，对缅甸语的音节划分和罗马化进行了研究。研究依据缅语文字的书写规范和音节结构，归纳了缅语音节边界的划分规则，并最终实现了缅语音节的自动划分。同时，依据缅语的音位系统和 MLC 转写系统，制订并实现了与其读音相对应的缅甸语文本罗马化方案。实验结果表明，此种音节划分和罗马化方法，具有唯一性，可满足缅甸语文语转换系统的要求。

10. 缅甸语 OCR 方法

缅甸语 OCR 方法和技术不成熟，已在某种程度上限制了缅甸语的自然语言处理、机器翻译和信息检索等问题的研究。为解决此问题，毛存礼等[258]提出了一种基于知识蒸馏的缅甸语 OCR 方法。该方法首次将基于知识蒸馏的思想运用到缅甸语图像文本识别研究中，构建了学生网络和教师网络对长序列中局部特征的增强，实现局部特征对齐，从而解决缅甸语嵌套组合字符识别的问题。实验结果表明，在没有背景噪声图像和有背景噪声图像作为训练数据集的情况下，这种模型的性能分别优于基线 2.9%和 2.7%。

11. 未来挑战和发展趋势

综合来看，缅甸语自然语言处理方面的研究，不论是数据资源建设、方法探索、技术开发，还是研究成果的转化应用等，仍处于初级阶段。现有研究，虽已涉及诸多领域，但大部分仍停留在较为基础的层面。部分领域问题的研究，如自动文摘、话题分析、文本纠错、预训练模型等，仍有诸多空白，有待进一步深入研究。这一方面是因为

缅甸语本身的一些特殊属性，导致很多技术和方法无法应用于缅甸语的加工处理，进而限制了相关研究深入和发展。另一方面，也与专门人才的缺乏密切相关。未来，应采取“自然语言处理人才+缅甸语专业人才”的模式，合力推进相关研究不断向前发展。

本章编写人员：

蒋盛益、张新猛、丘心颖、肖莉娴、陈诗、张卫国、武智

第 4 章 多语种语料与评测

4.1 引言

多语种智能信息处理技术评测是对多语种自然语言理解和多语种自动生成技术水平和系统能力的评估手段，是通过量化手段进行打分排序从而反映多语种智能信息系统性能优劣的过程。根据不同的分类方式，评测可分为白盒评测和黑盒评测、自动评测和人工评测、单项能力评测和综合能力评测等。1950 年提出的图灵测试被视作最早的自然语言处理任务的评测任务。当前针对评测任务的研究同时得到了研究界和工业界的广泛关注。

多语种智能信息处理评测目标是覆盖多语种、多领域智能信息处理任务，旨在推动通用的、鲁棒的多语种智能信息处理系统的研究，保持我国在多语种智能信息处理这一领域的传统优势和战略制高点，促进我国多语种信息处理技术和成果在国家“一带一路”建设中的辐射、引领性的作用，助力我国多语种信息处理技术健康、高速地发展，为多语种智能信息处理领域的繁荣和发展提供保障。随着多语种智能信息处理技术研究的迅速发展，针对评测的理论和方法研究就成了当务之急。

对于多语种智能信息处理的研究和开发者而言，公共评测通过给出量化结果来引导、支撑技术的公开对比和优化迭代。权威的评测活动是多语种智能信息处理研究的重要推动力，促进并引导研究人员面向实际应用需求，重视之前未被发现或有意回避的研究难点和重点。多语种智能信息处理公开评测任务的组织与实施，可以有效地促进多语种智能信息处理中共同性的、基础性的关键核心问题的解决。另一方面，对多语种智能信息处理系统的使用者而言，评测结果有助于帮助他们有效地在不用产品或不同系统间做出明智的选择。

4.2 多语种评测资源库建设现状

随着人工智能技术的不断进步与发展，多语种智能信息处理公共技术评测也在逐步开展。多语种智能信息处理基础资源及技术评测资源库建设也得到越来越多的关注和重视。从最开始的面向语言本体研究的言语资料集合到如今支撑自然语言处理的深度标注知识资源，语言资源库建设及相关研究深度和广度两方面得到了充分的探索。目前，国内外研究者无论是在用于词法、句法分析的语料库建设还是用于深度语义理解的语言知识库建设方面都做了大量的工作，并且取得了较好的成果。语言数据联盟（Linguistic Data Consortium，LDC）是具有影响力的全球语言资源平台，其包括语音和文本资源，涉及英语、汉语、阿拉伯语、波斯语等语言。目前 LDC 官网[1]罗列了 1993-2022 的所有语言资源库，其中最著名的句法树库 Penn Tree Bank 覆盖多种语言，包括中文（CTB）、OPEN MT 平行语料等涵盖了各种各样的语言资源。我国语言资源库的研究起步于上世纪 80 年代，近三、四十年来，语言资源建设的研究对象从单一的汉语语言资源库发展到多语种的语料库，其研究内容从面向语言学的研究拓展到人工智能等多领域的知识挖掘和知识发现。目前，国内也建立了中文语言资源联盟[2]（Chinese LDC），联盟已经拥有各类语音数据库、简体中文分词评测语料、CWMT 机器翻译测试语料、汉语情感语料库以及汉英、维汉综合领域平行语料库等资源库近七十余种。百度联合中国计算机协会和中国中文信息学会共同建立了面向自然语言理解和生成任务的“千言”（LUGE）中文开源数据平台[3]，涵盖了 10 大任务、36 个中文开源数据集，包括开放域对话、阅读理解、机器同传、情感分析、语义解析、信息抽取和文本相似度等语言资源库。除此之外，以清华大学、北大大学、北语语言大学、哈尔滨工业大学、科大讯飞，阿里

[1] https://catalog.ldc.upenn.edu/byyear

[2] http://www.chineseldc.org/

[3] https://luge.ai/

巴巴、国家语委、中科院、社科院为首的产学研界构建了丰富的语言资源库，具有代表性的语言资源库以汉语语料库为主，主要包括国家语委现代汉语通用平衡语料库（1 亿字）、北京语言大学语料库中心 BCC 语料库、清华汉语树库(Tshinghua Chinese Treebank，TCT)、北京大学 CCL 语料库、人民日报标注语料库，以及中国科学院、清华大学、南京大学构建的英汉、汉英双语平行语料库等。这些语言资源库的建设为计算机科学和语言学的共同发展搭建了重要纽带和桥梁。

目前，英、汉语等资源丰富语言的资源库建设取得了系列重大成果，为国内外语言学、计算机科学等领域的研究者提供了充足的数据支撑，但是，国内少数民族语言属低资源语言，其语言资源库建设起步晚，成果积累较少。国家历来重视少数民族语言文字信息化建设事业，制定了蒙、藏、维、哈、朝、彝、壮、傣、柯等少数民族文字编码字符集、键盘、字模等国家标准，研发了多种少数民族文字排版系统、智慧语音翻译系统等，支持少数民族语言文字网站和新兴传播媒体的有序发展，不断提升少数民族语言文字的信息化社会应用能力。近五年来，以中国科学院、中国社科院、中央民族大学、新疆大学、西北民族大学、青海师范大学、西藏大学、内蒙古大学等为代表的高校、科研机构及 IT 企业在民族语言多语种及蒙、藏、维单语种的信息处理及语言资源建设方面取得了令人可喜的成绩。由中央民族大学、清华大学、西藏大学组织承办了两次全国少数民族语言分词评测（Minority Language Word Segmentation）MLWS 2017 和 MLWS 2021 任务，两次任务积累的评测语料 MLWS 2021 已面向社会免费开放。MLWS2021 包含蒙、藏、维 3 个语种，数据规模扩大到目前的 15.5 万句的标注语料。评测任务是面向全国的蒙古文、藏文、维吾尔文三个语种的自动分词技术评测，推动了少数民族语言文字信息处理核心技术的交流与发展，以及民族语言开放资源的建设与共享。以下重点

阐述以蒙、藏、维为代表的少数民族语言资源建设研究进展情况。

4.2.1 蒙古语语料库建设

蒙古语语料库为蒙古语言学的研究提供了可靠的数据同时也为蒙古语的计算机处理研究提供了定量的有价值的信息，在蒙古语语言资源库建设方面内蒙古大学、内蒙古社会科学院、呼和浩特民族学院等研究单位做了很多的贡献。

1. 现代蒙古语语料库

内蒙古大学蒙古语研究所 1991 年建立了 100 万词级现代蒙古语语料库，1998 年扩充到了 500 万词语料库，涵盖了文科教材、理科教材、文学、新闻、政治、社会科学、自然科学、口语等类型语料。标注了词干词缀切分标记、复合词标记和人名、地名标记，其中 30 万词带有词性标注。

2. 蒙古文分词评测语料库

由中央民族大学、清华大学、西藏大学组织承办少数民族语言分词评测 MLWS2021 数据集在 MLWS2017 的基础上，由之前单一的新闻领域扩充到新闻、经济、法律、娱乐等综合领域。MLWS2021 数据集中的蒙古文由中央民族大学提供，共计 6.5 万句分词标注语料。

3. 汉蒙双语对照语料库

截止 2022 年，全国机器翻译大会（CCMT）已经连续举办了十七届，并组织了十一次机器翻译评测活动。全国机器翻译评测大会（CCMT）为了进行蒙汉机器翻译评测，开源了 25.6 万句对的平行语料库，推动了蒙古语机器翻译研究。

4.2.2 藏语语料库建设

藏文在浅层的字处理、词处理成果相对丰富，但是当上升到更深层次的句法、语义研究时，由于缺乏大规模的语言资源，导致研究进展缓慢，成果积累少。目前，藏语已经构建了藏语词法分析语料库、句法树库、双语平行语料库及一些语言知识库。

1.词法分析语料库

少数民族语言分词评测MLWS2021 数据集中的藏文由西藏大学提供，共计 2.5 万句；维吾尔文由清华大学提供，共计 6.5 万句。3 个语种的语料均来源于由新闻、经济、法律、娱乐等多领域组成的综合领域语料，语料来源于 2013 年西藏大学构建的“大型藏文基础语料库”，该语料库涵盖 1 亿 5 千万藏文字符的大型藏文平衡语料库，并对 3 千万藏文字符进行了分词和词性标注加工。此外，青海师范大学藏文信息处理与机器翻译省级重点实验室已完成 1000 万字的藏语语料库，用于藏语词语自动切分和标注；西北民族大学多拉团队构建了 500 万词次的藏文分词标注语料库；中国社会科学院民族学与人类学研究所建立的人工分词平衡语料，这些语言资源库为藏语自动分词研究提供了强大的数据支撑。

2. 句法树库

句法树库标注了十分丰富的词语形态信息、词类信息、句法结构信息、句法功能信息及语义角色信息。一个标注精细、合理的句法树库不仅可以供语言学家更好地研究语言的词汇、短语、句法等问题，也可为计算机处理自然语言提供优质的实验数据。目前中国社会科学院民族学与人类学研究所龙从军等建立了 1.0 万句基本句型的藏语短语结构树库；华却才让等以半自动的方式构建了 1.1 万句藏语依存树库，扎西加、多拉等构建了 1 万句藏语依存树库；夏吾吉等人工构建了 2106 句藏语语义依存树库。

3. 语义知识库

清华大学张钹院士在 2017 年提出“AI 未来的科学突破是建立一种同时基于知识和数据的 AI 系统。”语义知识库包含了人类认知的本体知识、语言知识、自然知识等基础知识体系，是计算机从“弱人工智能”转为“强人工智能”的关键所在。

语义知识库的构建是目前自然语言深度语义理解的重要基础，尤

其在低资源语言研究中，语义知识库能够有效提高模型的性能。藏语语义知识库构建工作基本是 2000 年开始的，且大都参考了汉语或英语的相关成果，例如多杰卓玛引入框架理论（FrameNet）来分析藏语语义。将同一认知场景中的词纳入一个框架之中，以此对藏语进行知识上的表述；才让三智等更具藏语传统语言文法构建了藏语虚词知识库；祁坤钰参照 WordNet 提出了藏语语义词典设计理念，用同义词集合表示概念；柔特以 WordNet 为基础，利用自动映射的方法对齐 WordNet 词典数据库和藏语数据库中的共有词汇，然后用人工的方式对词汇缺省问题进行处理；龙从军等根据哈工大的同义词词林构建了藏语语义相似词知识库。姚洲从语义角色出发，提出了藏文 Hownet 语义知识库构建方法。总的来说，目前藏语语义知识库的构建层次、数量和规模仍十分匮乏。

4. 机器翻译双语评测语料

藏语机器翻译的双语语料也基本来源于 CCMT 开源的评测语料。目前该评测集的藏汉双语翻译语料规模达到 15. 7 万平行句对，为藏汉机器翻译的发展提供了数据。

5. 其他

除了上述语言资源建设外，藏语自然语言处理相关研究单位还构建了用于文本分类、情感分析的语言资源库，如复旦大学开源了藏文文本分类评测语言资源库，该数据集包括 52,131 个不同类别的文本，平均每个文本包括 689 个音节，文本的标题平均为 16 个音节；西藏大学杨欣、群诺等构建了 4723 条句子的藏文情感语料库，该语料库情感类别包括 8 大类和 21 个小类。

4.2.3 维哈柯语语料库建设

新疆大学多语种信息技术重点实验室、新疆民族语音语言信息处理实验室、新疆师范大学等研究单位为维吾尔语的语言信息处理做了大量的工作。

1. 词法分析语料库

维吾尔语在语料库建设方面已做了大量的工作。新疆大学吐尔根·依布拉音等和新疆师范大学的玉素甫·艾白都拉等构建了百万词次的维吾尔语词法分析语料库，并分别在这些语料库基础上进行了词法、句法及面向具体任务的标注。维吾尔语的词语形态切分是将一个词切分成形态或语素的结构化预测任务，是自然语言处理领域基础且重要的内容，其输出结果能够帮助提高各种不同应用任务的性能。清华大学哈里旦木·阿布都克里木等构建了维吾尔语形态切分语料库(THUUyMorph)，该语料库包含 10596 个文档、69200 个句子，词语类型为 89923 个，分为词级和句子级两类标注，开源网址为（http://thuuymorph. thunlp. org/）；此外，Kahaerjiang 等建立了小规模命名实体关系语料库。

少数民族语言分词评测 MLWS2021 数据集中维吾尔文由清华大学提供，共计 6.5 万句，来源于由新闻、经济、法律、娱乐等多领域组成的综合领域语料。

2. 双语平行语料库

新疆大学的新疆多语种信息技术重点实验室公开了维吾尔语-汉语综合领域平行语料库(http://www. chineseldc. org/)，该语料库包含 5 万维吾尔语-汉语句对；阿西穆·托合提等构建了乌兹别克语-维吾尔语双语语料库构建平台，该平台已经构建了包含 8124 条句对的双语对齐语料库；冯韬等研究并设计了汉维可比语料库，目前该语料库包含 5000 个汉维可比语料篇章，主要是新闻领域语料和政府公文等。此外，全国机器翻译评测大会（CCMT）为了进行维汉机器翻译评测，开源了 17 万句对的平行语料库。

3. 语义知识库

四川电子科技大学阿里甫. 库尔班参考英语 FrameNet，结合维吾尔语源语言的框架语义描述体系，构建了词一级的维吾尔语框架语义

知识库（UFN），目前 UFN 已就维吾尔语名词、形容词、动词、量词和副词等 4252 个词元构建了 402 个框架，其中 2700 个词元完成了例句标注，总共标注了 1.85 万例句的框架语义信息。新疆大学吾买尔江·库尔班构建了以配价作为基本描写法、真实语料为事实依据的维吾尔语框架语义知识库（简称框架网 Frame Net），该知识库在构建维吾尔语词汇及其所属框架的语义词典等诸多领域有着广阔的应用空间和发展前景

4. 其他

目前，维吾尔语的情感分析语料库方面的研究相对成熟，如新疆师范大学年梅等构建了包含了将近 4500 个褒贬情感词的情感词库；新疆大学伊尔夏提·吐尔贡等构建了 11261 条维吾尔文情感语料库，该语料库制定了 8 个大的情感类别和 25 个更加细致的情感类别；此外，新疆大学阿布都热依木·热合曼通过研究国内外相关的句法树库标注体系建设理论，再结合维吾尔语自身的特点，制定了维吾尔语句法树库标注体系规范，采用了人工标注与自动标注相结合的方式完成了 3000 句规模的维吾尔语句法树库，为今后维吾尔句法树库研究的不断深入奠定了一定的基础。

4.2.4 其他低资源语言

少数民族语言的低资源属性，限制了语言信息化进程，目前除了蒙、藏、维（哈柯）三种少数民族语言资源语言，国内研究机构还根据不同需求，构建了一些其他少数民族语言资源库。例如西南民族大学民族语言文字信息处理技术研发中心 2010 年构建了彝汉双语词语标注语料库、彝汉人名汉字音译数据库、彝族传统医药术语数据库；2012 年构建了西南民族大学王成平采用彝族经典创世史诗——《勒俄特衣》为语料，构建了信息处理用彝、汉、英三语平行语料库。此外，周秀苗构建了库容为 69 万词的壮族典籍多语平行语料库；张羽构建了壮、汉、英三语平行语料库等。

4.3 多语种评测技术与研究现状

4.3.1 多语种分词评测

1. 中文分词评测

中文分词已有 30 余年的研究历史，是中文自然语言处理任务的基础与核心，其研究成果被应用到自然语言处理的不同任务中，包括信息检索、机器翻译、语音识别、文本错误识别、中文繁简体自动转换、自动问答等。中文分词首先要有清晰的分词标准，然而中文博大精深，分词标准一直以来都无法统一。目前，只能对具体问题设定特定标准。在特定标准下，实际分词的过程中还存在切分歧义和未登录词识别两大问题。

SIGHAN 是国际计算语言学协会中文处理特别兴趣组，自 2003 起共举行 9 次研讨会，发表中文分词相关文献达 76 篇。另外 SIGHAN 和中国中文信息学会（CIPS）先后三次举办中文处理资源与评测国际会议（CIPS-SIGHAN），不断推动着中文分词技术的发展。

SIGHAN 采用多家机构的评测数据组织多次评测（即 BakeOff），评测使用封闭测试（Close）和开放测试（Open）两种方法。封闭测试只允许使用固定训练语料学习相应的模型，而开放测试可以使用任意资源。测试使用的评价指标包括准确率、召回率和 F 值。其中，对比的黄金标准是人工标注的数据集。它们至今仍作为学术界评测分词方法准确程度的重要标准。表 4-1 及表 4-2 显示 2003 年至 2010 年，BakeOff 各个语料在封闭测试和开放测试上得分最高的队伍、F1 值和测试语料中 OOV 的比率。表 4-3 显示了 CIPS-SIGHAN2012/2014 两届的最好评测结果。

由评测结果可知，2003 年至今，分词方法从基于“词+规则”慢慢转变成基于机器学习字序列标注的方法。字序列标注的方法也经历了：单独使用一个机器学习模型、机器学习模型+简单的人工/训练语

料统计信息后处理、机器学习模型+无监督测试语料统计信息后处理、多个同类机器学习模型投票、多种机器学习复杂模型分层处理，这样一个演变过程。机器学习方法所使用的特征也越来越有效。近三年来，深度学习、LSTM、双向 LSTM 和注意力机制是中文分词研究的主流方法。综合比赛评测和文献，各种中文分词技术在 SIGHAN 数据集上评测的最佳 F 值基本达到甚至超过 95%。单纯设计一种学习算法已很难继续提升分词精度，如何更有效地结合不同算法是未来的研究方向。

表 4-1 BakeOff 评测结果（2003-2010）

年份	任务		F1	OOV 率	年份	任务		F1	OOV 率
	类型	语料				类型	语料		
2003	Close	AS	0.961	0.022	2005	Close	AS	0.952	0.043
		CTB	0.881	0.181			PK	0.950	0.058
		HK	0.940	0.071			CityU	0.943	0.074
		PK	0.951	0.069			MSR	0.964	0.026
	Open	AS	0.904	0.022		Open	AS	0.956	0.043
		CTB	0.912	0.181			PK	0.969	0.058
		HK	0.956	0.071			CityU	0.962	0.074
		PK	0.959	0.069			MSR	0.972	0.026
2006	Close	CITYU	0.972	0.040	2010	Close	L	0.946	0.069
		CKIP	0.958	0.042			C	0.951	0.152
		MSRA	0.963	0.034			M	0.939	0.110
		UPUC	0.9333	0.088			F	0.959	0.087
	Open	CITYU	0.977	0.040		Open	L	0.955	0.069
		CKIP	0.959	0.042			C	0.950	0.152
		MSRA	0.979	0.034			M	0.938	0.110
		UPUC	0.944	0.088			F	0.960	0.087

表 4-2 BakeOff 评测结果（2007）

年份	语料	Close		Open	
		F1	OOV 率	F1	OOV 率
2007	CITYU	0.851	0.082	0.969	0.082
	CKIP	0.947	0.074	0.956	0.074
	CTB	0.959	0.055	0.992	0.055
	NCC	0.940	0.047	0.975	0.047
	SXU	0.962	0.051	0.973	0.051

表 4-3 BakeOff 最佳系统评测结果（2012/2014）

	Precision	Recall	F-Measure
2012	0.946	0.950	0.948
2014	0.968	0.978	0.973

2. 蒙藏维分词评测

少数民族语言文本的分词处理和中文分词一样是语言信息处理的基础性工作，是民族语言机器翻译、智能检索、自然语言理解与处理等智能信息应用的前提。少数民族语言分词技术评测(MLWS)由中国中文信息学会主导，中央民族大学、清华大学、西藏大学联合发起，自 2017 年始已成功举办两届，评测对象是蒙古文、维吾尔文、藏文三个语种的自动分词核心技术，评测简介见下表。相较于第一届，第二届在评测语料规模和领域上进行了大幅度升级：领域由之前单一的新闻领域扩充到新闻、经济、法律、娱乐等多领域；数据规模也由之前的 3 万多句，扩大到目前的 15.5 万句，其中蒙古文 6.5 万句，维吾尔文 6.5 万句，藏文 2.5 万句。评测使用正确率、召回率、F 值评价各个参评单位的分词结果。

表 4-4 少数民族语言分词评测项目表

序号	项目代号	项目名称	语种	训练语料	测试语料	语料领域
1	MO	蒙古文分词	蒙古文	6.5w	2w	综合领域
2	UY	维吾尔文分词	维吾尔文	6.5w	2w	综合领域
3	TI	藏文分词	藏文	2.5w	2w	综合领域

评测吸引了来自民族信息处理领域的众多研究者踊跃报名，以藏文分词评测为例，两届藏文分词评测的最终结果见下表。可知，藏文的分词结果无论在准确率、召回率还是 F 值均有很大的提升，少数民

族语言分词技术有了长足的进步。

表 4-5 MLWS2017 藏文文本分词评测结果（新闻领域）

序号	参赛代码	版本	准确率 P（%）	召回率 R（%）	F 值（%）
1	T4	contrast-b	93.34	92.46	92.90
2	T6	primary	93.32	91.82	92.56
3	T3	primary	91.04	91.62	91.33

4.3.2 机器翻译评测

机器翻译评测是对给定翻译系统生成译文的质量进行量化的评价。最早的机器翻译评测可以追溯到 1964 年，当时美国国家科学院的语言自动处理咨询委员会（ALPAC）通过人工的方式对译文质量进行评测。机器翻译评测与机器翻译技术的进步相辅相成，其价值不仅仅在于评价机器翻译质量，还能够及时给机器翻译研究人员反馈机器翻译本身存在的问题，指导其如何改进及优化。

国际机器翻译大赛（WMT）是全球学术界公认的国际顶级机器翻译比赛。自 2006 年至今，WMT 已成功举办 16 届，每次比赛都是全球各大高校、科技公司与学术机构展示自身机器翻译实力的平台，更见证了机器翻译技术的不断进步。其它国际赛事还有美国国家标准和技术机构(NIST)组织的机器翻译比赛和语音语言技术国际研讨会(IWSLT)协办的文本翻译赛事、IWSLT 口语机器翻译评测等，地区性的赛事包括中国机器翻译研讨会(CWMT)。

到现在为止，机器翻译评测已经形成比较完整的分类体系。通常机器翻译评测可以分为人工评测与自动评测，下面分别阐述。

1. 人工评测

人工评测是指由人通过主观判断对机器译文进行打分。目前常用的人工打分指标包括流利度、忠实度、理解力、清晰度和保真度。其中，忠实度、流利度指标出现最早，应用最为广泛。表 4-6 为国际口

语翻译评测中对忠实度与流利度的等级划分情况。除此之外，针对特定的应用场合，人工评测还会使用其它评测指标，如基于任务、后编辑、子集排序等。

表 4-6 忠实度与流利度的等级划分情况

等级	忠实度	等级	流利度
5	传达全部原文的意思	5	流畅的英语
4	传达大部分原文的意思	4	较好的英语
3	传达较多原文的意思	3	非母语表达的英语
2	传达较少原文的意思	2	不流畅的英语
1	没有传达出原文的意思	1	表达不清的英语

人是机器翻译质量的最终评定者，但是人工评测也存在耗时、昂贵等问题。此外，在很多情况下，人的评价标准可能因为评价者及评测时间不同等主客观因素而发生变化，使得评测结果不可重复，因此自动评测方法成为技术和实践上的双重需求。

2. 自动评测

根据评测过程是否依赖参考译文，自动评测又分为基于参考译文的自动评测和译文质量评估两种类型。基于参考译文的自动评测模型，通过计算自动译文输出和参考译文之间的相似度来评价翻译质量。不依赖参考译文的评价模型大多依赖机器学习的特征，从源语言的原句子和目标语言的译文里提取有效特征来估计译文质量，这些特征可以包括词性、句法、语言模型等。与人工评价相比，自动评测的好处包括廉价、快速、可重复性、可用来调整和优化机器翻译的模型参数等。自动评测的常用方法包括以下几类：

编辑距离。编辑距离指将一个字符串转化为另一个字符串所需的最少编辑操作次数，许可的编辑操作包括插入、删除、替换。编辑距离是一种经典的相似度计算方法，编辑距离越小说明两个字符串越相似。常见的基于编辑距离的自动评测指标有 WER、PER、TER、mWER、

mPER、mTER 等。

n 元匹配。用机器译文中出现的 n 元文法与参考译文中出现的 n 元文法进行比较，计算完全匹配的 n 元文法的个数与机器译文中 n 元文法的个数的比值。这是一种类似准确率的计算方法，它允许一个原文有多个参考译文。常见的基于 n 元匹配的自动评测方法有 BLEU、NIST 等。BLEU 是目前使用最广泛的自动评测方法，是用于评估模型生成的句子和实际句子差异的指标。它的取值范围在 0.0 到 1.0 之间，如果两个句子完美匹配，BLEU 是 1.0；反之，如果两个句子完全不匹配，BLEU 为 0.0。

语言学特征。基于语言学的自动评测引入了一些语言学资源来解决 n 元文法纯字面匹配的缺陷。这些语言学特征包括侧重于词性、词组、句子结构的句法信息以及侧重于同义词、语义等的语义学信息。典型的基于语言学的评测方法有 METEOR、WoodPecker 等。

深度学习。早期基于神经网络的机器翻译评测工作在结构上使用简单的前馈神经网络，在学习特征上使用基于词向量的句子级别语义信息，在机器学习模型的设计上使用成对比较的方法，分别计算两个不同机器翻译的输出与参考翻译的相近程度来选出较好的机器翻译。随着深度学习技术的发展，长短期记忆神经网络、注意力机制、预训练模型等新的技术逐渐加入并诞生了一系列优秀的评测方法。典型的评测方法包括 BERTscore、BLEURT、COMET 等。

4.3.3 语音识别评测

语音识别技术应用领域众多，语音识别系统的性能评测对语音识别技术的发展起着重要的推动作用。CHiME 为国际多通道语音分离和识别大赛，由法国计算机科学与自动化研究所、英国谢菲尔德大学、美国三菱电子研究实验室等知名研究机构所于 2011 年发起，至今已举办 6 届，是业界影响力最大、参赛队伍最多、水平最高的多通道噪声鲁棒性语音识别比赛。此外，东方语种识别国际竞赛 OLR Challenge

将目标聚焦在东亚（中国、日本、韩国等）以及东南亚（印尼、越南等）地区的多语种语音识别研究，目前已举办 6 届，评测语种由最初 2016 年的 7 种扩大到 2021 年的 18 种，训练数据更长达 180 小时之多。

为探索低资源条件下的语音识别技术，自 2020 年起，美国国家标准与技术研究院 NIST 连续发起 OpenASR20 和 OpenASR21。该赛事设置的主要目的是在多语种语音识别任务上探索如何使用少量的数据达到较好的效果，同时考察低资源语音识别基础算法在多个语种上的推广性。以 OpenASR21 为例，比赛共包含 15 个语种，见下表。涵盖受限赛道、受限附加赛道和非受限赛道。其中受限赛道为各参赛单位必选项，每个语种只能使用组委会提供的 10 小时标注语音识别数据，受限附加赛道在受限赛道的基础上允许使用开源的预训练模型，而非受限赛道可以使用组委会提供 10 小时受限数据之外的数据。

表 4-7 openASR21 比赛语种

语言	使用国家	语言	使用国家	语言	使用国家
粤语	中国等	普什图语	阿富汗等	他加禄语	菲律宾等
瓜拉尼语	巴拉圭等	索马里语	索马里	格鲁吉亚语	格鲁吉亚等
爪哇语	印度尼西亚等	泰米尔语	印度等	哈萨克语	哈萨克斯坦等
库尔德语	伊拉克等	越南语	越南等	阿姆哈拉语	埃塞俄比亚等
蒙古语	蒙古国等	斯瓦希里语	坦桑尼亚等	波斯语	伊朗等

4.3.4 其它

通用语言理解评估基准（GLUE）是自然语言处理领域公认的最具权威的语言理解评测基准之一，由来自纽约大学、华盛顿大学、Google DeepMind 等机构的研究者在 2018 年共同推出，用于评估和分析多种已有自然语言理解任务的模型性能。GLUE 包含九项自然语

言理解任务，语言均为英语，涉及到自然语言推断、文本蕴含、情感分析、语义相似等。之后，很多模型在 GLUE 的大部分任务上都达到了 90 分以上，GLUE 基准在新模型的评估能力渐渐达到上限。随后 SuperGLUE 应运而生，并凭借多样化任务、全方位的考察能力受到产学界的广泛追捧。截至 2020 年年底，SuperGLUE 排行榜上效果最好的模型 T5 已经非常接近人类水平。虽然 GLUE 基准升级，但它针对的是英文任务。ChineseGLUE(简称 CLUE)是中文版本的多任务自然语言理解基准与分析平台，共包含 10 个中文语言评测任务，其中很多任务所用的数据集是 GLUE 数据集的中文版。2021 年 12 月 30 日，机器中文语言能力评测基准——智源指数 CUGE 发布，它涵盖 7 种重要语言能力、17 个主流任务、19 个代表性数据集，是兼顾自然语言理解（NLU）与自然语言生成（NLG）两大任务体系的中文语言能力评测标准。类似地，2019 年法语综合评测基准 FLUE、2020 年印尼语评测基准 IndoNLU 相继提出。

GLUE、CLUE、CUGE 等综合能力评测仅限于单语种，无法对低资源语言作出相应评测。为考察模型零样本跨语言迁移学习能力，2020 年 3 月，来自 CMU、谷歌研究院和 DeepMind 的科学家们提出覆盖 40 多种语言（跨 12 个语系）的大规模多语言多任务基准 XTREME；2020 年 5 月，微软发布 XGLUE 基准数据集，由 11 种任务组成，涵盖 19 种语言。与 XTREME 相比，XGLUE 可同时评估跨语言预训练模型在跨语言自然语言理解和生成方面的性能。

除了上述多任务评测基准，也有很多单项能力评测基准，此处仅介绍几类经典的单项能力评测及其相关的数据集。

表 4-8 经典任务及数据集

序号	任务	数据集	描述
1	命名实体识别	CoNLL2003	该任务给定一个句子，要求机器正确识别句子中人名、地名等包含名称的短语。CoNLL2003 数据集包括 1393 篇英语新闻文章和 909 篇德

序号	任务	数据集	描述
			语新闻文章。
2	关系推理	TACRED	TACRED 是一个拥有 106264 条实例的大规模关系抽取数据集，这些数据来自于每年的 TACKBP 比赛使用的语料库中的新闻专线和网络文本。
3	阅读理解	SQuAD	在原来的 SQuAD 的 10 万个问题答案对的基础上，SQuAD 2.0 中新增了超过 5 万个由众包工作者对抗性地设计的无法回答的问题。
4	情感分类	SST	SST 属于单个句子的文本分类任务。输入一个句子，要求输出该句子的情感倾向，即输出“非常积极”“积极”“中立”“消极”或“非常消极”中任一类型。SST 包含 11855 个句子及相应情感标签，更有挑战性的是，SST 同时给出了这些句子的语法分析树中 215154 个短语的细粒度情感标签。
5	文本摘要	DUC	文本摘要任务给定一段长文本，要求机器输出保留其主要信息的短句。DUC 系列中最被广泛使用的是 DUC2004 数据集，其包含 500 组文档摘要对，文档平均 35.6 个词，摘要平均 10.4 个词，

4.4 产学研应用及行业技术评测发展现状

4.4.1 行业技术评测工作概况

行业技术评测主要指由企业或行业联盟所发起的，针对特定行业领域的技术应用评测。从企业参与评测组织的角度而言，主要可分为两大类：一是在各类学术会议的竞赛单元中，企业以联合举办的形式开展各类垂直领域技术评测，二是由企业独立发布的技术评测任务。

第一类主要体现在包括在中国计算语言学大会（CCL）、全国知识图谱与语义计算大会（CCKS）、全国社会媒体处理大会（SMP）、CCF 大数据与计算智能大赛（CCF BDCI）等会议中，均有企业与大会主办方联合设置的针对具体领域的技术评测任务。学术会议中涉及行业的具体技术评测任务中，金融、医疗、法律、电商是关注较多的领域，CCKS 从 2021 年开始关注军事领域的技术评测。各类会议结合自身关注的学术领域，从不同角度开展面向具体应用的评测任务。

企业独立发布的技术评测任务也可具体细分为两种形式：一是企业搭建技术评测平台，开展常规滚动评测的同时，配合具体需要发布具体评测任务。该类平台包括由阿里巴巴搭建的阿里云天池[4]、由百度搭建的飞桨[5]、爱奇艺搭建的爱奇艺 AI 竞赛平台[6]、华为搭建的华为云平台[7]、科大讯飞搭建的讯飞开放平台[8]等。除爱奇艺 AI 竞赛平台外，其余平台发布的评测任务均不局限于企业主营业务，阿里云天池、飞桨等平台所发布的评测任务甚至不仅包括行业技术应用评测，还进行了系列针对算法的学术类技术评测。而为了更好地吸引学术界参与行业技术评测，上述平台除讯飞开放平台外，均有与国际、国内顶会合作发布任务的意愿和偏好。二是由行业或企业主办的专项技术评测，包括中国法律智能技术评测（CAIL）、中国数字人文开放创新研究大赛、中国健康信息处理会议（CHIP）等，该类专项技术评测通常领域更加聚焦，具备一定的规划性和整体性。

4.4.2 行业技术评测领域分析

从近 5 年行业技术评测任务关注领域来看，除传统关注较多的金融、医疗、法律、电商、新闻、工业、（汽车、电力、航空等）等领域外，文化艺术、军事、教育等领域也在逐步成为行业关注的热点，以下将选取主要领域进行分析。

1. 金融

金融领域是自然语言处理技术最早进行应用落地的领域之一，主要应用集中在个人或机构风险识别、智能客服问答、产品或平台用户评价提取、事件抽取、金融知识图谱构建等方面。近两年关注重点逐渐向小样本迁移学习、篇章级事件抽取和事件因果关系的抽取、知识图谱的自动构建等方向转移。

[4] https://tianchi.aliyun.com/

[5] https://aistudio.baidu.com/aistudio/competition

[6] http://challenge.ai.iqiyi.com/

[7] https://competition.huaweicloud.com/home

[8] http://challenge.xfyun.cn/

2. 医疗

医疗领域的自然语言处理技术应用主要集中在面向临床的应用，包括中文病例实体识别和关系抽取、医疗知识图谱构建、医疗对话的理解与生成、临床术语标准化，以及面向药物研发的蛋白质结构预测、药物与分子链接预测等。伴随中医应用的不断扩展，面向中医临床和中医文献的相应评测任务也在逐步开展。同时，伴随互联网医疗的兴起，传统面向临床医疗的任务逐渐从电子病例等相对标准化结构文本转向非标准化文本（如病人自述、互联网医患对话等）的实体识别、关系抽取和事件抽取，医疗对话生成和电子病例生成类任务也受到关注。

3. 法律

在法律领域，国内较为权威且和系统的专业领域评测任务是中国法律智能技术评测（CAIL），从 2018 年开始至今，已连续举办 4 届。每年发布评测任务都有所差异，其中较为稳定的任务包括阅读理解、类案检索、司法考试、司法摘要等。此外，在 CCF BDCI、SMP 等会议中也设置有法律领域评测任务。纵观近 5 年领域内评测任务，其变化一方面体现在任务领域的不断细分和扩增，如论辩理解、信息抽取、事件检测等新任务的出现；另一方面也体现在传统任务的升级上，如阅读理解任务的文书种类由民事向刑事、民事、行政的扩增、问题类型由单步预测转向多步推理，司法摘要由单文档摘要转为多文档摘要等。

4. 文化艺术

文化艺术领域是近年来新兴的人工智能应用场景，较为集中的应用于文献、古籍的挖掘与展示，包括文本挖掘、社会关系分析、知识图谱构建以及古籍整理（句读、实体识别、词性标注等）等，也出现了用于文言文处理的大规模预训练模型；少量分散分布于影视剧、音乐、游戏等领域，运用自然语言处理等技术进行情感分析、意图理解、

赛况预测、平台用户留存度分析等任务处理。总体来看，目前在文化艺术领域的评测任务处于起步阶段，应用技术的范围和深度都相对局限，进一步深入研究提升空间较大。

5. 军事与网络安全

伴随国家安全问题关注度的日益增加，军事领域相关任务评测开始受到学术界和行业关注，主要在 CCKS 会议竞赛单元中有所涉及。2021 年在知识图谱构建与问答主题下设置了面向军用无人机系统的军事垂直领域知识图谱构建的子任务；2022 年单独设置军事知识图谱主题，包含两个任务子项。网络安全领域中，主要应用于反诈和互联网金融安全，包括金融负面信息及主体判断、广告欺诈判定等。

4.4.3 行业技术评测资源库建设

行业资源库建设方面，目前国内建设相对较少，多为根据特定评测任务所构建的评测数据集。仅在医疗、法律等领域有较为成熟的多任务评测数据集资源。同时，在中文通用领域，百度联合高校力量，构建了千言数据集。数据集内目前汇集了 14 所高校和企业的 36 个开源数据集，并针对 10 个具体任务开展相应评测。以下将主要针对多任务数据集进行介绍。

1. 金融领域资源库

SmoothNLP 金融领域文本数据集：该数据集包含约 50 万条企业工商信息、210 万金融讯息新闻、58 万专栏资讯、3 万投资机构信息、7 万投资时间以及 11 万 36 氪新闻，可用于词嵌入、实体识别、无监督聚类、企业行业分类、标题总结、文本分类等任务训练。

2. 医疗领域资源库

（1）CBLUE：中文医疗信息处理评测基准 CBLUE（Chinese Biomedical Language Understanding Evaluation）是中国中文信息学会医疗健康与生物信息处理专业委员会在合法开放共享的理念下发起，由阿里云天池平台承办，并由医渡云（北京）技术有限公司、平安医

疗科技、阿里夸克、腾讯天衍实验室、北京大学、鹏城实验室、哈尔滨工业大学（深圳）、郑州大学、同济大学、中山大学、复旦大学等开展智慧医疗研究的单位共同协办。目前已发布至2.0版本，在1.0版本的基础上丰富了语料来源，扩充了任务类型。目前主要包括医学文本信息抽取（实体识别、关系抽取、事件抽取）、医学术语标准化、医学文本分类、医学句子语义关系判定、医学对话理解与生成共5大类任务14个子任务。

（2）cMedQA2：cMedQA2 是一个中文医疗问答数据集，目前发布至2.0版本，包含10万余条问题数据和20万余条答案数据，可用于医疗领域对话理解与生成等领域任务训练。

3. 法律领域资源库

（1）中文法律阅读理解数据集CJRC：中文法律阅读理解数据集CJRC是由哈工大讯飞联合实验室提出，是收个中文法律阅读理解数据集，包含约10万篇文档，最终形成约5万个问答对。主要涉及民事一审判决书和刑事一审判决书，包含约188种民事案由、138种刑事罪名，数据来源于中国裁判文书网。该数据集涉及问题类型众多，可应用于要素抽取、信息检索、问答系统等任务训练。

（2）无标注数据库：法律领域有众多无标注的数据库，收集数据范围涵盖法律法规、司法案例、裁判文书、法学期刊等多种类型，影响力较大的有北大法宝法律数据库、北大法意——中国法律资源全互动数据库、中国裁判文书网、无讼案例、威科先行法律数据库等。

4. 文化艺术领域资源库

（1）C-CLUE：C-CLUE 由天津大学贡献，是一个基于众包标注系统构建的文言文语言理解测评基准及数据集，基于二十四史构建。目前开源了由系统标注结果获取的近2万个实体以及4千多个关系，可供自然语言处理中命名实体识别和关系抽取任务直接使用。

（2）古文现代文翻译平行语料库：该数据库基本涵盖大部分经

典古籍著作，包括《论语》、《孟子》、《左转》、《资治通鉴》等短篇章，以及二十四史、《太平广记》、《徐霞客游记》、《水经注》等典籍，共计形成文言文-现代汉语平行句对约 96 万。

（3）殆知阁古代文献集：殆知阁古代文献数据集包含佛藏、儒藏、医藏、史藏、子藏、易藏、艺藏、诗藏、道藏、集藏等十大类约 20 万卷古代文献，均为 TXT 版本，共计约 20 亿字。是古文领域基础语料，可用于古文分词、句读、词性标注、情感分析等任务。古文领域预训练模型 GuwenBERT 即是根据此数据集训练而来。

此外，企业搭建的评测平台如阿里云天池、百度飞桨等，以及部分行业平台如 openKG[9]、Digital Humanities Portal[10]、和鲸社区[11]等也搭建了专区用以汇聚和展示行业开源数据集，所收集数据集既包括本平台组织评测所用数据集，也包括领域内使用范围较广的经典数据集。

总体而言，国内的行业技术评测参与主体范围逐步扩大，参与形式也在不断更新和改进。但由行业或企业主办的专项评测还不够成体系、不够规范，而由学会主办的评测任务又无法聚焦于某一具体领域进行深入挖掘，因此总体存在整体性不强、连续性较弱的现象。从领域角度来看，行业评测关注领域在不断拓展，传统领域的评测任务则更加细化和深入，多模态技术、事理图谱构建、半监督/无监督学习、大规模预训练模型等相关技术正在广泛应用于各垂直领域具体任务当中。在数据资源建设方面，目前国内针对具体行业领域数据集建设已初见成效，在金融、法律、医疗等较为成熟的自然语言处理应用领域形成了一批数据标注清晰、应用范围广泛的资源集；但仍存在数据规模小、缺乏统一标注规范等问题。需要构建领域内统一格式、能够对模型进行多维度综合评价的数据集。

[9] http://www.openkg.cn/

[10] https://www.dhlib.cn/

[11] https://www.heywhale.com/

4.5 展望

评测对多语种信息处理研究具有重大指导意义，目前针对评测的研究吸引了学术界和产业界的广泛兴趣。现阶段在多语种信息处理的各类评测任务的组织和实施过程中也面临很多困难，包括缺少评测所需的多语种评测数据集尤其是大规模评测数据集，缺少通用的高质量的评测标准，缺乏评测普适的评测原则和评测规范等。专委会今后的研究热点是构建面向多语种信息处理任务的高质量评测数据集，并制定高质量评价标准和方法，为多语种信息处理研究提供客观、公平、公正、开放的评测结果，不断引领和推动多语种智能处理技术的研究和行业应用的各类模型、方法的创新和进步。

本章编写人员：

赵小兵、李琳、高璐、高歌、周毛克

第 5 章 多语种预训练语言模型

自然语言理解是自然语言处理任务中最难的一部分，本章从自然语言任务出发，首先介绍自然语言理解的过程，然后介绍当前在自然语言处理任务中最先进的预训练语言模型。

5.1 自然语言理解的感知与认知

5.1.1 从感知到认知

人工智能（AI）是跨越计算机科学、数学、认知科学及神经科学等学科的一门先进技术。自 1956 年，在美国达特茅斯会议上正式提出 AI 之后，其发展主要经历了三个时期。

20 世纪 90 年代以前，采用专家系统和知识工程的方法，构建“知识 + 逻辑符号”系统来模拟人类的智能阶段，称为知识（规则）驱动的 AI；然而，受限于当时人工知识（规则）对自然语言的描述能力，这一时期 AI 基本局限在实验室研究范畴。

从 20 世纪 90 年代中期直到近几年，AI 的机器学习相继跨入到统计机器学习及深度机器学习时期，称之为数据驱动的 AI 时代。这一阶段由于机器性能的大幅提升，以大规模真实语言数据训练自然语言处理（NLP）模型成为可能，并逐步广泛进入社会应用领域。然而，上述方法的缺陷在于，第一代知识驱动的 AI 主要靠人工从原始数据中获取知识，效率低、规则描述能力有限等；第二代数据驱动的 AI 可以从训练数据中自主地获取知识，但其性能受到数据规模和质量限制，鲁棒性差，易受干扰，是“黑箱操作”。

为了建立一个全面反映人类智能的 AI，需要建立鲁棒性强、可解释的 AI 理论与方法，即第三代 AI。2018 年底，张钹院士公开提出第三代“知识 + 数据”双轮驱动 AI 的理论框架体系。因此，在基于大数据的深度学习进入发展的“瓶颈”期后，从 2018 年至今，AI 开始进入到发展的第三个时期。这一时期不再只关注数据，知识的获取重新得到了极大重视。实现真正的智能系统，需要将数据和知识进行

深度融合，在数据上要有归纳能力，能够举十得一；在知识上，要有逻辑推理能力，能够举一反三。

虽然，当前基于大数据驱动的深度学习方法，能够挖掘高维数据复杂的结构特征，并用人类熟悉的方式沟通和互动，具备了视觉、听觉和触觉等感知能力，在语音、图像、文本和视频识别等方面已经逐渐接近甚至超越了人类的水平。但是，在数据驱动的感知 AI 框架中，只要轻微变动图像、文本或语音数据就可以欺骗这些已经训练好的系统，造成感知误判。

人类引以为傲的认知能力，都是以语言为载体进行的。自然语言理解（NLU），是第三代 AI 的终极目标，旨在赋予机器阅读和理解人类语言的能力。由于人类自然语言的复杂性，目前的机器学习系统仅能进行数据处理，并不能真正理解数据的含义，通过缩小任务范围或扩大数据集来回避处理语义的问题，机器只是“记录”数据，但没有“理解”数据，所以机器在 NLU 方面的表现远不如人类。

5.1.2 自然语言理解的难点

自然语言总是涉及对现实世界事件的描述。实现对自然语言的理解，需要依赖人类常识及上下文语境，挖掘语言潜在语义的逻辑和因果关系。由于自然语言本体的一些固有特性，也会导致计算机语言理解的困难。人类理解和生成语言，依赖词汇、句法、语义等语言本体知识，以及自然常识、人文和自然科学知识等。对于机器来说，基本要求是具备一定的逻辑推理能力和认知能力。认知活动最本质的特点是利用知识来指导行为，涉及三个方面的内容，首先是信息的获取、表示并转化为机器知识；其次是知识的存储和提取；最后是运用知识进行推理等处理过程。认知过程主要是知识存储并利用知识进行语义推导。为使计算机具备一定的认知能力，需要对各类知识进行形式化表示，以及用能够让计算机可以识别的形式加以合理地描述和存贮。因此，实现真正的 NLU 需要解决两个问题，首先获取、表示及计算

隐含的、高度多样化的多源知识；其次，整合这些抽象知识到 AI 系统中，帮助机器进行语义理解和常识推理。

在第一代和第二代 AI 的发展过程中，NLP 的两大代表性方法为基于知识的方法和基于统计的方法。

1. 基于知识的方法

专家系统和知识工程作为认知智能的早期代表，学者们提出“将知识引入 AI 领域”，为计算机理解自然语言建造了各种知识库，此类研究一般以某种语言为主体语言设计知识库的框架结构，并以此为基础添加其他语言。目前，项目开发成熟、较有影响力的语言知识库有 WordNet、FrameNet、PropBank、HowNet 等。

通过知识库系统确定句子中每个单词的作用，并提取上下文的含义。知识库提供了良好的逻辑性和可解释性的语言分析方法，但却严重依赖人工定义的范畴与规则。虽然人类是用其全部的经验与知识来理解和生成语言的，但是人工知识库仍然难以完整地表示人类的经验和知识并全部编码进入计算机，这类知识缺少对特征抽象和学习的能力。

2. 基于统计的方法

受限于人工知识库存在规模较小、自动构建能力不足、知识获取困难等一系列问题，学界出现了从大量数据的概率分布中学习基于统计的模型和方法。为了让计算机处理语言文本，需要将字、词、段落等信息转换为机器可以理解的方式进行，以便在计算机中表示语言或文本，并能让计算机程序自动处理，这就是语言表示。

早期的语言表示是以词袋模型、N 元模型为代表的离散表示，仅仅将词符号化，词与词之间没有距离的概念，两个词只要字面不同就难以刻画它们之间的联系，比如“电脑”和“计算机”这样的同义词会被看成是两个不同词。因此，导致语义鸿沟、维度灾难等问题的出现。与离散表示不同的是连续表示，将语言表示为连续空间中的一个

点，即连续向量。这种表示的优势可以把对文本内容处理简化为连续向量空间中向量运算，通过计算向量空间上的相似度，来表示文本语义上的相似度，计算机很容易处理“向量”，因此取得很好效果。但是，将文本以向量形式表示时，忽略了词语内部语义或词序信息的考量，也出现了不少问题。如代表性的 Word2vec 方法给出的在向量空间相似度较高的词并不会考虑语义，仅仅以在句子中的用法是否相同作为衡量标准。且 Word2Vec 基于上下文学习方法的词向量技术倾向于把贡献较多的词语聚在一起，可以学习到上下文语境相似的词汇，却难以捕获到深层词汇语义的相似性，特别是语料中出现频次较低的词语义项的相似性。

5.1.3 语言知识图谱

近年来，深度学习技术充分利用神经网络的分布式表示能力和层次结构泛化能力，从大规模数据中自动学习，显著提升了对无结构文本、图像、语音数据背后语义信息的表示与学习性能，将数据驱动方法推向新高度。另一方面，纯数据驱动深度学习是输入和输出之间的特征关系，不具备因果推理性，缺少可解释性。对大规模数据的学习与利用，离不开深度学习技术，但要实现有理解能力的 AI，还需要人类认知知识作为支撑。

语言知识图谱是实现认知智能的解决方法之一，不同于传统知识工程的“小知识”，以知识图谱为代表的大数据时代各种知识系统，受益于海量数据、强大算力、最优算法，能够自动构建大规模、多领域、高质量的知识库，形成所谓的“大知识”。知识图谱把非结构化、离散的知识以图结构形式组织起来，从而描述关于世界万物的概念、实体、事件及其之间的关系。知识图谱包含的背景，赋予机器精准查询、深度理解与逻辑推理等能力，被广泛运用于实体消歧、推荐系统、问答系统和复杂问题推理等任务，在认知智能实现中起到非常重要的作用。知识图谱与深度学习方法相结合，一方面深度学习可以从数据

中（有标注数据、弱标注数据及无标注数据）学习和挖掘有用信息，为大规模知识图谱的补全提供支持；另一方面，知识图谱技术的成熟，获取的知识也可以被用于深度学习的知识指导，为知识融入深度学习框架提供了理论基础。

知识图谱根据所含知识类型的不同，可大致分为三种。语言知识图谱。自然语言具备的词法、句法、语义、语篇及语用等方面的语言知识，如 WordNet、HowNet 是典型的词法知识图谱。常识知识图谱。人类对自然界事物普遍认知的日常共识知识，如 Cyc、ConceptNet 是典型的常识知识图谱。社会知识图谱。现实世界中人类社会活动产生的各实体之间的事实知识和关系，如 WikiDatas、Freebase、DBpedia、YAGO 是典型的社会知识图谱。除这些有典型代表的知识图谱外，还有涵盖特定专业及业务领域的专业知识图谱及商业知识图谱。以上知识类型划分并没有严格界限，如 HowNet 既包含词法级语言知识，也包含大量的常识知识。

5.2 预训练语言模型

BERT(Bidirectional Encoder Representations from Transformers)的出现彻底改变了自然语言处理领域，并在各种任务上取得了最先进的性能。其训练范式为采用大量的单语数据训练一个预训练模型，然后采用少量特定任务的数据对模型进行微调。预训练使用掩码语言建模目标进行，从而学习到更好的语义表示。鉴于 BERT 模型在英语自然语言处理中的成功应用，许多语言特定的 BERT 不断涌现，例如法语的 FlauBERT、荷兰语的 BERTje、芬兰语的 FinBERT 等。然而，语言资源的规模为这种特定语言模型的训练带来了一定的挑战与限制。

5.2.1 预训练语言模型介绍

预训练属于迁移学习的范畴。现有的神经网络在进行训练时，一般基于后向传播（Back Propagation，BP）算法，先对网络中的参数

进行随机初始化，再利用随机梯度下降（Stochastic Gradient Descent，SGD）等优化算法不断优化模型参数。而预训练的思想是，模型参数不再是随机初始化的，而是通过一些任务进行预先训练，得到一套模型参数，然后用这套参数对模型进行初始化与训练。

5.2.1.1 预训练模型架构

多语种预训练模型的目标是学习一个可以生成给定文本的多语言表示的模型，该模型可以在公共向量空间中为跨语言的相似句子和单词生成相似的表示。典型的多语种预训练模型包括 N 层 Transformer 的编码器(Encoder)，每层包含 k 个注意力头(Attention Head)，与一个前馈神经网络(Feedforward Neural Network)。对于输入序列中的每个标记，注意力头使用句子中所有其他标记表示的注意力加权线性组合来计算嵌入，然后将来自所有注意力头的嵌入连接起来，并通过前馈网络为每个输入标记生成一个 d 维的嵌入表示。现有的多语种预训练模型可能在 N、k 和 d 的选择上有所不同。最后一个编码器层的输出通常用作每个标记的上下文表示，而与[CLS]标记对应的嵌入被认为是整个输入文本的嵌入。

5.2.1.2 预训练任务

各种多语种预训练模型在模型架构上基本大同小异，其主要差别在于设置的预训练任务与损失函数上。根据所需的训练数据的不同，这些预训练任务可以大致分为单语预训练任务或多语预训练任务。

1. 单语预训练任务

单语预训练任务仅在单语数据上开展，这些任务的目标是无监督或者自监督，通过在给定上下文标记的情况下预测缺失标记来训练模型生成多语言语义表示。包含掩码语言建模、因果语言建模和因果语言建模。掩码语言建模是用于训练大多数多语种预训练模型的预训练任务之一。通常，其他预训练任务与掩码语言建模结合使用。它是将来自多种语言的单语数据汇集在一起，将单一语言的无监督掩码语言

建模简单地扩展到多种语言。对于给定训练示例中的单词序列，掩码语言建模任务随机选择 15%的标记进行掩码，对于待掩码的标记，80%的概率替换成[MASK]标记，10%的概率替换成随机标记，10%的概率保持原标记，目标是使用剩余的标记来预测这些被掩码的标记，即该模型被训练以最小化用于预测掩码标记的交叉熵损失。因果语言建模是传统的语言建模目标，即在给定先前标记的情况下预测下一个标记。与掩码语言建模不同，因果语言建模只能获取到待预测单词前面的信息。对于给定训练示例中的单词序列，因果语言建模的目标是在给定前 i-1 个标记的情况下预测第 i 个词，即该模型被训练以最小化在给定前 i-1 个标记的情况下预测第 i 个标记的交叉熵损失。

2. 多语种预训练任务

多语预训练任务旨在明确强制跨语言的相似文本表示在多语言语义空间中彼此接近，目标是单词级别或句子级别。由于平行语料库通常比单语语料库小得多，因此多语预训练任务常常与单语预训练任务结合使用，通过多语预训练任务和单语预训练任务的联合优化来完成。多语种预训练任务主要有翻译语言建模、跨注意力掩码语言建模、跨语种掩码语言建模、跨语言对比学习、跨语言句子对齐和翻译替换标记检测。

翻译语言建模是指给定语种 A 与语种 B 的平行文本序列，两个序列都作为掩码语言建模的输入，采用[SEP]标记将两个序列拼接起来。与掩码语言建模类似，翻译语言建模对部分标记进行掩码这些标记要么属于语种 A 中的序列，要么属于语种 B 中的序列。为了预测 A 中的被掩码的标记，模型可以依赖语种 A 中的上下文信息或语种 B 中的翻译信息，从而隐式地被迫学习语种对齐的表示。更具体地说，如果语种 A 中的上下文表示不充分，则模型可以使用语种 B 中的上下文表示来预测语种 A 中的被掩码的标记。翻译语言建模的最终目标函数与掩码语言建模相同，即最小化掩码标记的交叉熵损失，唯一

的区别是被屏蔽的标记可以属于任何一种语言。

跨注意力掩码语言建模通过预测平行句子对中的掩码标记来学习跨语言表示。在预测源句子中的掩码标记时，该模型仅限于使用目标句子的语义信息，反之亦然。与翻译语言建模不同，在翻译语言建模中，模型可以访问两个输入句子来预测掩码标记，在跨注意力掩码语言建模中，模型被限制为仅使用相应平行句子中的标记来预测源句子中的掩码标记。

跨语种掩码语言建模与翻译语言建模目标非常相似，都是使用跨语言句子作为输入，其主要区别在于，跨语种掩码语言建模的输入是在文档级别构建的，其中来自跨语言文档的多个句子被替换为另一种语言的翻译。

跨语言对比学习的目标是最大化平行语料对中的句子之间的互信息度。由于在每一轮迭代过程中，模型不可能将训练数据中样本都计算一遍，因此跨语言对比学习使用了动量对比方法去构建对比样本，在这种方法中之前编码过的句子会被当作负样本复用。此外，对每个平行句子对来说，跨语言对比学习将其与另一个平行语料库中的随机平行句子对相拼接。这种数据增强能够鼓励预训练模型去判断多语言文本的次序和学习句子边界。

层次对比学习是跨语言对比学习的扩展，以学习句子级别和单词级别的跨语言表示。对于句子级损失，层次对比学习使用平滑线性插值在嵌入空间中的句子之间构建负样本。对于词级损失，层次对比学习使用句子表示和其他单词表示之间的相似度表示。对于每个平行句子，句子中存在的标记被认为是正样本，而词汇表中的其他标记则被认为是负样本。

跨语言句子对齐通过利用平行语料训练跨语言模型来对齐句子表示。给定一个平行句子对，该模型被训练以从小批量的负样本中预测句子 X 的相应翻译 Y。与对两个句子进行编码并使用[CLS]进行拼

接得到的句子表示不同，跨语言句子对齐的相应句子表示是通过对多语种预训练模型最后一层中的词嵌入进行平均来计算的。

类似于多语种替换标记检测任务，翻译替换标记检测任务在判别设置中利用平行的句子对，通过拼接平行句子以形成单个输入句子，然后使用生成器来预测掩码标记，将损坏的句子传递给鉴别器，鉴别器进行标记级别分类以预测是原始标记还是由生成器生成的标记。

5.2.1.3 经典的预训练语言模型

当前预训练语言模型主要有 BERT(Bidirectional Encoder Representations for Transformers)、RobBERTa(Robustly Optimized BERT Approach)、ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)等。BERT 旨在通过双向联合上下文进行表示建模，即从未标记的文本中学习深度双向文本表示。它由多个 Transformer 的编码端(Encoder)组成。如**图 5-1** 所示，BERT 的预训练阶段由掩码语言建模(Masked Language Model，MLM) 和下一句预测(Next Sentence Prediction，NSP) 两个无监督任务组成，MLM 是指从输入序列中屏蔽一些词，然后通过上下文预测被屏蔽的词；NSP 旨在增强句对之间的关系，它的目标是预测句子对是否连续。BERT 模型可以针对各种下游任务进行微调，例如文本分类、命名实体识别和自动问答任务等。

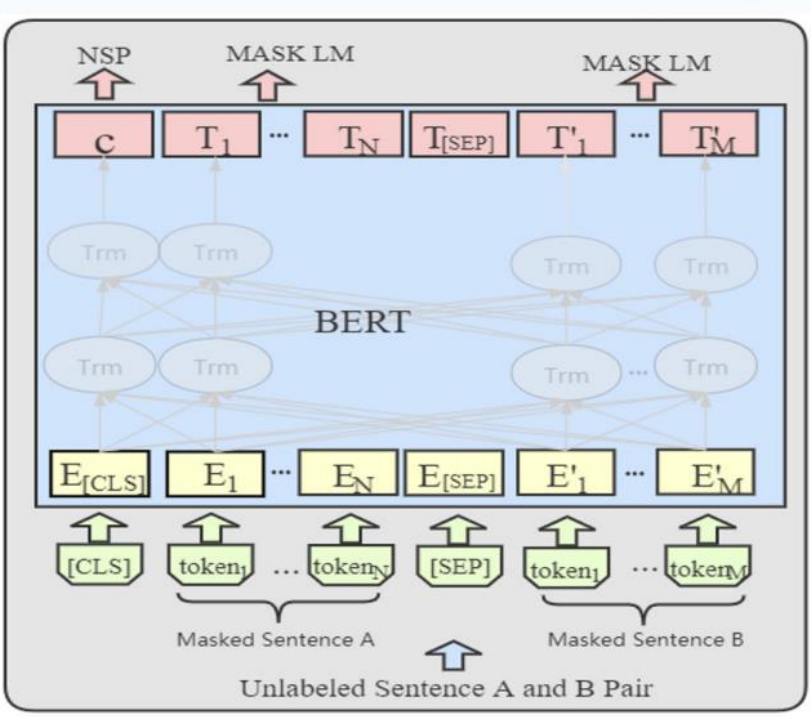


图 5-1 BERT 模型

作为 BERT 的变体，RoBERTa 旨在充分利用 BERT 架构和训练方法。与 BERT 相比，RoBERTa 有三个改进。(1)更多的训练数据。RoBERTa 利用更多的未标记数据对模型进行预训练，以在下游任务中获得更稳健的性能。(2)去除 NSP 任务。Liu 等验证了 NSP 任务的无效性并移除了该任务。(3)动态词掩码。RoBERTa 使用动态词掩码来优化 MLM 任务，而不是采用 BERT 模型提出的静态词掩码，可以让预训练模型的参数得到更充分的优化，从而模型可以更好地捕获序列特征。

图 5-2 展示了 ELECTRA 模型。与 BERT 相比，ELECTRA 的创新点如下：(1)提出了替换标记检测任务，该任务预测输入样本中的每个标记是否被生成器样本替换；(2)该模型联合训练一个小型生成器和一个判别器，以减轻判别器的训练难度；(3)为了有效地学习上下文信息，ELECTRA 使用权重共享的方式将生成器的 embedding 信息共享给判别器。

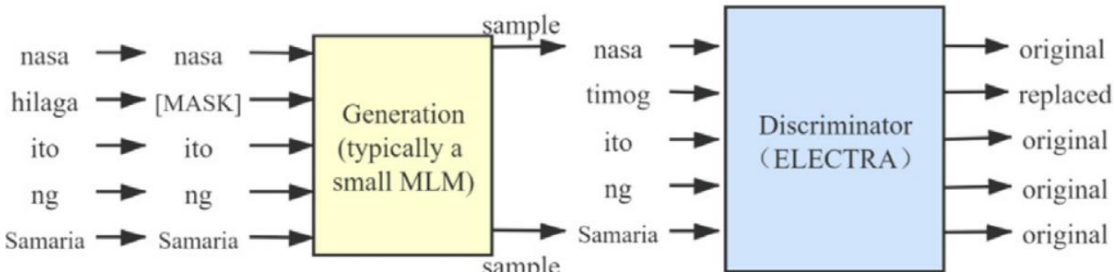


图 5-2 ELECTRA 模型

5.2.2 单语预训练模型

面向单语言的预训练模型研究主要集中在英语、汉语等富资源语言上，这对低资源语言来说，具有一定的限制与不良影响，如何将预训练技术更好地应用于各种低资源语言，成为学者们关注的问题。一种可行的方法是训练多语种预训练模型(Multilingual Language Models, MLLM)，而随着 mBERT(Muli-lingual Bidirectional Encoder Representations from Transformers)、XLM(Cross-lingual Language Model)、XLM-R 等多语言语言模型的出现，预训练模型也成为解决多语言与跨语言任务的重要方法。

多语种预训练模型使用来自多种语言的大量未标记数据进行预训练，希望通过共享词汇、遗传相关性(Genetic Relatedness)或联系相关性(Contact Relatedness)等方式构建不同语言之间的桥梁，从而使低资源语言可以从富资源语言中受益。因此，研究面向多语种预训练模型的工作具有广阔的前景。

以低资源东盟语种为例，当前针对低资源东盟语种的预训练模型的工作主要分为原始模型构建、基于语料规模扩充的模型构建、针对领域特殊性的模型构建、针对语言特殊性的模型构建，以及共享其他语言信息的模型构建五类。

1. 原始模型构建

作为预训练语言资源最丰富的东盟语种，马来语已经有了几十个

不同类型的预训练模型，涵盖 BERT、XLNet、AlBERT、T5、GPT2 等，这些模型均由 Malay Huggingface 构建，然而他们并没有公开各模型的训练细节与性能评测。而与马来语同源的印尼语，尽管有近 2 亿的使用人口，并且是世界上第十大使用语言，其预训练语言模型的关注度反而不高，与其使用地位显然不相称。由于缺乏带注释的数据集、语言资源的稀疏性和资源标准化，以前关于印尼语的工作受到了阻碍。Koto 等发布了 IndoBERT，一种面向印度尼西亚语预训练语言模型，并构建了 IndoLEM 印尼评估语料库，对该预训练模型进行基准测试。IndoBERT 与原始的 BERT 模型结构一样，Koto 等用大小为 31,923 的印尼语 WordPiece 词汇训练 IndoBERT，其训练语料超过 2.2 亿个 token，其来源主要有：①印度尼西亚语维基百科；②来自 Kompas、Tempo 和 Liputan 的新闻文本；③印度尼西亚网络语料库。Cruz 和 Cheng 使用 WikiTextTL-39 数据集预训练了一个菲律宾语 BERT 模型；此外，他们还通过模型蒸馏构建了一个较小版本的预训练模型 DistilBERT 模型。Cruz 等使用 WikiText TL-39 数据集训练了四个不同版本的 ELECTRA 模型。

2. 基于语料规模扩充的模型构建

在印尼语预训练模型方面，Wilie 构建了一个更大的印尼语语料库 Indo4B，该语料库共包含 250 万个句子、4 亿个 token，用于训练信息量更丰富的印尼语 BERT-BASE 模型和 BERT-LARGE 模型，模型的参数设置与原始的 BERT 模型一样，词汇表中共有 30,522 个 token。而对于菲律宾语预训练模型，Jiang 等采用更大的语料库(Oscar 语料库、维基百科语料库和新闻语料库)与更大的词表(52,000 个 token)训练了三个预训练模型，分别是 BERT、ELECTRA 和 RoBERTa。在预训练阶段，除了现有的开源语料库之外，他们还构建了一个大规模的新闻文本语料库用于预训练。结果显示，预训练语料更充足的模型性能更优异。

3. 针对领域特殊性的模型构建

在印尼语上，Koto 等提出了 IndoBERTweet，这是面向印尼语 Twitter 文本的第一个大规模预训练模型。该模型在 IndoBERT 的基础上，扩展附加领域特定词汇的词表进一步训练，同时关注词汇不匹配下的高效模型适应问题，并对新词初始化 BERT 嵌入层的不同方法进行基准测试，结果发现，使用平均 BERT 子词 embedding 进行初始化的方法可以使预训练速度快 5 倍。

4. 针对语言特殊性的模型构建

由于泰语、老挝语、缅甸语、柬埔寨语不像英语采用空格作为词与词之间的分隔符，因此在预处理时需要对输入的数据进行特殊处理。ThAIKeras 以泰语维基百科作为预训练语料，与原始的 BERT 切割方式 wordpiece 不同，他们采用 sentencepiec 作为文本的切割方式，训练了泰语 BERT 模型，其中 sentencepiece 模型采用 BPEmb 训练好的切分模型。Lin 等和 Jiang 等为老挝语、缅甸语提供了第一个基于 Transformer 的预训练语言模型，共包括 BERT-Small、BERT-Base、ELECTRA-Small 和 ELECTRA-Base 四个版本。在文本切割上，他们与 ThAIKeras 一样采用 sentencepiece 作为切分模型。柬埔寨语与老挝语、缅甸语、泰语相似，然而由于人们在编辑时习惯性加上空格使文本更加清晰，柬埔寨文本中存在大量空格去划分词语，因此与泰语、老挝语、缅甸语的预训练模型构建不同，Jiang 等在构建柬埔寨语的预训练模型时，没有采用 sentencepiece 模型，而采用与原始 BERT 一样的 wordpiece 算法。此外，他们尝试了分词与不分词两种策略，结果显示，即使存在分词算法的鲁棒性影响，先对文本进行分词的操作能提高预训练模型效果。

5. 共享其他语言信息的模型构建

Nguyen 等训练了一个越南语 RoBERTa 模型，所采用的数据来自网络媒体的新闻文本与维基百科语料库共 50 G；同时，还利用了引

入了英语的文本一起训练，从而使模型共享有其他语言信息，进而解决在越南语文本中存在英文单词的现象，使模型可以解决英语和越南语的语码混用现象。

5.2.3 多语预训练模型

面向多语言的预训练语言模型是处理多语言、跨语言任务的重要基石，同时也是处理低资源语言的重要基础。现有的多语言模型有Multilingual Bert(BERT)、Language-Agnostic Sentence Representations (LASER)、Language-agnostic BERT Sentence Embedding(LaBSE)、Crosslingual Language Model(XLM)等，处理的语种完全覆盖东盟国家所使用的语种模型仍是少数。

5.2.3.1 面向高资源语言的预训练模型

多语言预训练的突出优势在于能够将大量低成本收集、高成本标注的多语言训练数据汇聚在一起，经过大规模并行预训练去学习数据中的共性和语义，并把所有语言映射到统一语义空间中，然后将这些语义信息移植或者迁移到下游业务模型中，使得业务模型在少量标注数据上具备优异的泛化能力和鲁棒性能，甚至在某些语种上具备零样本学习能力。从这个角度出发，在高资源语言上，一般情况下数据规模越大、涵盖领域越丰富、数据异构性越强、数据质量越高会使多语言预训练模型能够学习到更鲁棒的语义信息。目前，诸如 mT5、DeltaLM、XLM-Roberta 等主流的多语言预训练模型，也验证了在高资源条件下，多语言模型的表现要媲美甚至优于单语言预训练模型。

高资源语言可以从三个方面来定义：大规模低成本未标注数据、小规模下游任务标注语料以及大规模多语言知识库。对于大规模低成本未标注数据资源，一般高资源语言全球使用人数较多或者使用范围较广，如中、英、俄、日、韩、德、法、西、葡、阿等，相应产生出的高质量人工编辑的资讯新闻数据、科技文献数据、百科数据较为丰富，如通用爬虫积累的多语言谷歌新闻资讯数据已达到几十 T 级的规

模，XLM-Roberta 采用了其中的 2.5T 多语言数据作为训练数据。对于小规模下游任务标注语料，由于统计自然语言发展已久，产生了大量的下游任务语料，如 ACE、MUC、WMT 以及其他公开的语料，这些也为高资源语言预训练提供了重要基础。对于大规模多语言知识库，目前知识和数据双驱动的预训练模型逐渐受到关注并得到了快速发展，高资源语言的知识信息能够为多语言预训练带来显著的增强学习能力。从目前多语言预训练模型的综合发展来看，相对而言，中、英、俄、日、韩、德、法、西、葡、阿等几十种语言能够达到高资源的规模，高资源数据建设任重而道远。

对于目前主流的面向高资源语言的多语言预训练模型，一般采用跟相对应的单语预训练语言模型相同的预训练任务和模型结构，不同的是多语言模型的词汇量会更大，比如单语种 BERT-BASE 模型的词汇量为 28,996，而多语言 BERT-BASE 基础模型词汇量则达到 119,547。mT5 采用了和 T5 相似的模型结构，并在 GeGLU 非线性、更大模型中缩放 dmodel 和无 dropout 等方面进行改进，mT5 采用了 mC4 数据集，覆盖了 100 多种语言。Facebook AI 团队于 2019 年 11 月发布了 XLM-RoBERTa 多语言训练模型，作为其在原始 XLM-100 模型的改进，所用训练数据达到 2.5T，涵盖 100 个语种，该模型在中、英等高资源多语言和单语言对比评测中都取得了出色的性能。另外，近两年诞生的比较出色的多语言预训练还有百度的 ERNIE-M、微软的 DeltaLM、字节跳动的 mRASP 等，都对高资源预训练模型的发展产生了深远影响。

高资源多语言预训练模型虽然取得了一定的进展，但是目前多语言预训练整体性能与单语种相比优势还是不明显，尤其是在高、中、低资源混杂的场景下，高资源对低资源的干扰还是较强，使得相比单语种性能存在降低的可能。针对此，高资源多语言预训练模型未来发展的方向主要有：（1）如何利用通用爬虫或者数据增强的方式，定向

的提高低资源语种的规模和丰富度;(2)如何改进多语言预训练模型,能够使得高资源语言减少对低资源语言的干扰,并如何建立高资源和低资源语言之间的共性关系,使得高资源更好的迁移提高低资源语言性能;(3)如何将多语言或者跨语言知识信息和数据进行相互增强或者驱动,利用各个语言的语法结构、百科知识和专家经验规则来指导预训练的大规模学习。

5.2.3.2 面向低资源语言的预训练模型

除了面向高资源语言,现有的多语言预训练模型也考虑了部分低资源语言。谷歌提出了多语言 BERT 模型,多语言 BERT 以与单语 BERT 相同的方式进行预训练,但它不是仅在英语单语数据上进行训练,而是在 104 种语言的维基百科语料上训练,并使用基于 WordPiece 切分模型的 119,547 个多语言共享 token。该多语言 BERT 模型覆盖了东盟 8 个语种中的马来语、印尼语、泰语、菲律宾语、缅甸语和越南语,由于老挝语和柬埔寨语的资源与使用人数少,因此没有在多语言 BERT 模型的训练语言中。

多语言 BERT 存在的主要缺点是在文本蕴含任务中,当前提和假设使用不同语言时,多语言 BERT 性能急剧下降。一种可能解释为 BERT 的学习方式是通过将前提中的单词或短语,与假设中的单词或短语进行匹配作出文本蕴含决策。LASER 模型对此进行改进,它支持文本推理任务中不同语言的前提和假设的任意组合;LASER 对所有输入语言使用一个共享编码器,并使用一个共享解码器来生成输出语言,模型输出的向量表示将任何语言的句子映射到高维空间中的一个点,目标是任何语言的相同语句都将出现在同一个邻域中。这种表示可以被视为语义向量空间中的通用语言,该空间中的距离与句子的语义接近度非常相关。XLM 采用两种学习跨语言语言模型的方法,一种是无监督学习,只依赖于单语言数据;另一种是监督学习,在平行语料数据上利用一个新的跨语言语言模型目标函数。所有语种共用

一个字典，该字典是通过 Byte Pair Encoding(BPE)构建，共享的内容包括相同的字母、符号 token(如数字符号)、专有名词。这种共享字典能够显著地提升不同语种在嵌入空间的对齐效果。XLM 不仅保留了 BERT 模型的 MLM，还采用因果语言建模(Causal Language Modeling，CLM)，在给定前序词语的情况下预测下一个词的概率，同时提出了翻译语言建模(Translation Language Modeling，TLM)将并行的翻译句子拼接起来，在 source 句子和 target 句子中都随机掩码掉部分 token，从而引导模型将两种语言的表征进行对齐。在多语言预训练模型中，尽管在进行 MLM 和 TLM 时学习到的内部模型表示形式对下游任务进行微调很有帮助，但它们不能直接产生句子嵌入，而这对于翻译任务至关重要。谷歌提出了 LaBSE 的多语言 BERT 嵌入模型，该模型使用 MLM 和 TLM 在 170 亿个单语句子和 60 亿个双语句子对上进行了训练。此外，LaBSE 还在翻译排名任务(Translation Ranking Task，TRT)上进行微调。TRT 使用带有共享变压器的双编码器体系结构进行训练，通过给定源语言中的句子，让模型进行排序，从而对目标语言中的句子的正确翻译进行排名，使多语模型在多项并行文本检索任务表现出最先进性能。

针对东盟 8 个官方语种的单语言预训练模型结构仍较为基础，未充分利用东盟语种的特性进行改进与优化。而多语种预训练语言模型中，采用的低资源语料较少，在低资源语种上的表现效果较差，如 Jiang 等在菲律宾语任务上证明了构建的单语言模型性能比 XLM 更优异，而 Lin 等则验证了 XLM 模型在老挝语上表现效果不佳。面向低资源语言的预训练模型仍然存在很大的不足，未来的主要研究方向有：(1)如何设计性能更优，分词更完整且词表更小的分词方法做为多语言的文本切分工具；(2)如何充分利用文本本身包含的词、短语等信息；(3)如何利用语言之间的相似性，使多个语言可以之间的信息可以相互利用。

5.2.4 多语种预训练模型的研究前景

现有的多语种预训练模型研究主要集中在四个方面的工作：(1)构建规模更大的多语种预训练模型以涵盖更多的语言；(2)创建更全面的基准以涵盖面向多语种预训练模型的更广泛的任务和更多的语言；(3)深入分析多语种预训练模型在单语、零样本跨语言和双语任务上的表现；(4)理解多语种预训练模型所包含的通用语言模式；(5)增强多语种预训练模型能力来提高其在可见甚至不可见语言上的性能。多语种预训练模型仍具有很大的探索空间与研究空间：

1. 零样本评估

多语种预训练模型的主要目标仍然是跨语言性能，尤其是零样本学习。然而，零样本学习的结果在不同任务和语言的已发表文献中存在很大差异，需要进行更系统的研究。

2. 语言包容性

多语种预训练模型有望成为世界上众多语言的“基础设施”资源。许多语言被广泛使用，但在研究和开发方面不够集中。为此，多语种预训练模型必须变得更具包容性，扩展到更多种语言。这可能需要模型创新，例如超越特定语言的适配器。至关重要的是，它还需要在各种任务和语言中提供包容性基准。

3. 模型的高效性

多语种预训练模型代表了目前最大的一些模型。但是，在边缘设备上通常无法在如此大的模型上运行推理，并且在云设备上的成本越来越高。一个重要的研究方向是在不影响准确性的情况下缩小这些大型模型。剪枝、量化、分解、蒸馏和架构搜索等几种标准方法已用于单语种模型。需要为多语种预训练模型探索这些方法，同时确保保留模型的通用性。

4. 模型的健壮性

支持多种语言的多语种预训练模型需要针对任何编码偏差及其

泛化能力进行广泛评估。研究方向之一是构建广泛的诊断和评估套件，还需要为一系列任务和语言开发评估框架，以识别多语言模型所犯错误的性质。

本章编写人员：

赵小兵、申影利、姚洲、蒋盛益、林楠铠、王连喜、张宝华、陈自岩

第 6 章 多语种词法分析

本章对基于神经网络自然语言处理方法中的子词切分方法进行了综述。首先解释基于神经网络自然语言处理方法中面临的，由于封闭词表所导致的集外词问题，并介绍了解决这一问题的 BPE、WordPiece 和 Unigram 这三种常见方法。子词切分之前通常需要做词语切分，而词语切分与具体语言高度相关。SentencePiece 提供了一种与语言无关的子词切分方法，可以在输入的句子上直接做子词切分。随后介绍了解决子词切分有时会存在一些切分不合理和子词表示学习不充分问题的子词正则化技术和 BPE-Dropout 技术；以及解决基于字符的子词切分在面对多语言（特别是中日韩等语言）的大字符集时依然存在 OOV 问题的一种有效手段——基于 UTF-8 字节的 BBPE 及其衍生的基于 BBPE 的 SentencePiece 方案；并对 ACL2021 最佳论文提出的一种通用的词表最优化技术 VOLT 进行了介绍。最后，介绍了多语种新词发现算法以及形态切分方法。

6.1 多语种词法分析概述

6.1.1 封闭词表假设和集外词问题

自然语言处理（NLP）所面临的文本词表（Vocabulary），通常是开放的，并没有任何规定文本中不能出现某些词。但作为一个 NLP 系统，通常都要假设所处理的词限定在一个有限的词表范围内，这个假设称为封闭词表假设（Closed Vocabulary Assumption）[297]。尤其是对于基于神经网络的 NLP 方法，由于我们需要把每个词事先映射到一个词嵌入表示（Embedding），而对于预定义词表以外的词（又称集外词；或者 Out-of-Vocabulary Word，简称为 OOV）是无法预知其嵌入表示的，因此也无法处理。这就是 NLP 系统所面临的集外词问题（简称 OOV 问题）。

6.1.2 集外词替换为 UNK

早期基于神经网络的 NLP 系统，对这一问题通常采用一个简单

粗暴的方法，即将所有集外词统一替换为 UNK，意为未知词（Unknown Word）。但这种做法显然不完美，原因有以下二点。

（1）很多集外词虽然在训练数据中没有出现过，但其意义并不是完全不可知；相反，相当多集外词的意义可以从已知词推导出来，把它们都替换成 UNK，就无法利用这些词的语义信息。比如：

◆ 屈折变化词。对某些词，词表中可能没有收录其所有的变化形式（如英语名词复数、形容词比较级和最高级、动词现在分词和过去分词等），这些没有被收录的就会被识别为集外词，如词集中可能有 cascade，但没有 cascaded。

◆ 复合词。比如 hard-working、thirteen-years 等。

◆ 数词。数词是不可能全收的，没有收录的数词都会成为集外词。

（2）在自然语言理解（NLU）任务中，出现 UNK 的问题对系统性能的影响不会很严重，但在自然语言生成（NLG）任务中，如机器翻译、对话生成等任务中，生成的句子中如果出现很多 UNK 是很严重的问题，用户会无法接受。简单删掉这些 UNK 又会使得句子不通顺。

6.1.3 基于字符的模型

基于字符的模型把模型建立在字符基础上，词的表示（Word Embedding）通过字符的表示（Character Embedding）经过一个神经网络计算得到。基于字符的模型可以较好地解决 OOV 问题，但也会带来序列长度大大增加，而模型处理长序列难度比处理短序列大得多。通常为了达到类似性能，基于字符的模型需要比基于词的模型复杂得多。

而对于中日韩（CJK）语言这样的大字符集语言，基于字符的模型需要的词表依然太大。Unicode 1.0.1（1992）定义的 CJK 统一字符集已经有 20902 个字符，而最新的 Unicode 14.0（2021）已经包含字

符总数 144697 个，把这么多的字符全部收入模型是不合理的。而且绝大部分字符都是罕用字符，使用概率极低，全部收入词表会极大地增加模型参数。不仅如此，由于每次生成一个字符需要在整个词表上做 softmax 操作，如此大的词表，且词表中大部分都是罕见字符，也会导致模型推理效率大大降低。

6.2 BPE 和 WordPiece

6.2.1 子词级别的切分

如前所述，词级别的词表无法解决集外词 OOV 问题，而字符级别的词表又存在模型复杂度高和效率低的问题。为此，学者们提出了子词（Subword）级别的词表构造和切分方法，可以较好地解决集外词问题。子词级切分方案的基本思想是，词表只收录固定的子词集合；任何一个词，如果不在词表中就需要被切分成子词序列（subword token sequence）。这种方法的优点是：彻底解决了集外词 OOV 问题，任何一个单词都可以切分成子词序列，最坏情况下，所有子词都匹配不上就会切分成字符序列；大部分常用词都可以作为独立子词收入词表，这样原始的基于词的模型无需做任何改动，序列长度增加很少，模型复杂度和效率都几乎没有增加。

问题是如何确定基本的子词词表——子词词表构造问题；如果确定子词词表，如何将句子切分成子词序列——子词切分（Subword Tokenization）问题。

对于大部分西方语言，一种很容易想到的方法是将词语切分成词根和词缀，词根和词缀就是子词。但是这种办法有以下问题：不是所有词都可以切分成词根和词缀的，比如外来词和数词等；词语切分成词根词缀的过程，通常被称为形态分析或者词法分析，而这个过程是和语言高度相关的，无法做到语言通用性。

目前，最流行的两种构造子词词表和子词切分的方法分别是

Byte-Pair Encoding（BPE）和 WordPiece，还有一种使用较少的 Unigram 方法，下面分别进行介绍。

6.2.2 BPE

BPE 最早是作为一种压缩算法提出来的[298]，Sennrich 等首次提出将 BPE 算法用于子词级切分，以解决神经机器翻译的集外词问题，效果非常明显。很快这一方法也被用于解决其他问题，并成为基于神经网络的 NLP 模型中最为流行的子词切分方法。

采用 BPE 算法做词例切分，首先需要使用语料库构造一个子词词表。下面以表 6-1 和表 6-2 为例将构造子词词表的过程介绍如下。表中，</w>是一个特殊字符，用于表示一个单词的结尾。step0 是初始化的词典和子词词表，其他每个 step 将出现频率最高 sub word pair 合并成一个新的 subword，并加在子词词表末尾。

表 6-1 BPE 子词切分过程示例——Dictionary

freq	step0	step1	step2	step3	step4	step5
5	low</w>	low</w>	low</w>	low</w>	low</w>	low</w>
2	lower</w>	lower</w>	lower</w>	lower</w>	lower</w>	lower</w>
6	newest</w>	newest</w>	newest</w>	newest</w>	newest</w>	newest</w>
3	widest</w>	widest</w>	widest</w>	widest</w>	widest</w>	widest</w>

步骤 1 预处理文本得到一部词典（Dictionary），词典中包含文本中出现的所有单词，以及该单词出现的频率。

步骤 2 将词典中所有单词切分成基本字符的序列，字符和字符之间加一个空格。

步骤 3 初始化子词词表（Vocabulary），词表只包含所有的基本字符（如字母、数字、标点、汉字等）构成子词 subword。

步骤 4 重复以下过程，每步将子词词表中的两个子词合并（merge）

得到一个新的子词加到词表末尾，直到子词 Vocabulary 的长度达到预设值：①词典中选择出现频率最高的子词对（subword pair）；②将上述子词对合并成一个新子词，并到子词 Vocabulary 末尾；③在词典中出现的所有上述子词对都合并成新加入子词（删掉原来两个子词之间的空格）。

步骤 5 返回词表。

表 6-2 BPE 子词切分过程示例——Vocabulary

freq	step0	step1	step2	step3	step4	step5
7	l	l	l	l	l	l
7	o	o	o	o	o	o
13	w	w	w	w	w	w
16	</w>	</w>	</w>	</w>	</w>	</w>
17	e	e	e	e	e	e
2	r	r	r	r	r	r
6	n	n	n	n	n	n
9	s	s	s	s	s	s
9	t	t	t	t	t	t
3	i	i	i	i	i	i
3	d	d	d	d	d	d
9		e+s	e+s	e+s	e+s	e+s
9			es+t	es+t	es+t	es+t
9				est+</w>	est+</w>	est+</w>
7					l +o	l +o
7						lo+w

基于 BPE 的子词切分非常简单，对于一个给定的文本，首先把

单词之间的空白都替换成</w>，然后把所有单词都切分成字符（任何两个字符直接都加一个空格），根据 BPE 的子词词表从前往后，对于每个 merge 形成的子词，在文本中反复执行其对应的合并操作，直到所有合并操作都执行完毕。

6.2.3 WordPiece

Google 最早在 Japanese and Korean Voice Search 一文中介绍了 WordPiece 的思想，但并没有命名这个算法。在论文 Google's Neural Machine Translation System : Bridging the Gap Between Human and Machine Translation 中首次提到在神经机器翻译系统中使用了 WordPiece 做文本的子词切分。后来 Google 在 BERT 中再次使用了 WordPiece，使得 WordPiece 影响大增。但 Google 并没有开源 WordPiece 的词表构造算法，也没有公开源代码。好在 BERT 的代码中还是包括了 WordPiece 的词例切分工具（Tokenizer），所以如果研究者们不想构造自己的子词词表，还是可以直接使用 BERT 提供的词表和词例切分工具做研究的。后来 Huggingface 的工具包中也提供了一个 WordPiece 模型训练工具，但他们并没有真正使用 Japanese and Korean Voice Search 文中所提算法，还是使用 BPE 方法构造了一个子词词表，只是把生成的词表改成了 BERT 提供的 WordPiece 数据格式。

WordPiece 的子词词表构造算法与 BPE 非常相似，也是一个从基本字符集开始逐渐合并的过程，每次合并都在词表中增加一个新子词。区别在于，BPE 在合并两个子词时，只需计算两个子词连续出现概率，而 WordPiece 则复杂得多。WordPiece 首先需要根据给定的子词词表构造一个 n-gram 语言模型，然后用这个语言模型可以计算整个训练数据出现的似然率（likelihood）。合并时，考虑所有的子词对，根据该子词对合并后得到新的语言模型计算整个训练数据的似然率，选择导致整个训练数据似然率最高的那个子词对进行合并。这个过程比较复杂，计算代价很高。Google 估计是设计了某种优化算法才有可能

实现高效的词表构建，但这个算法没有公开。推理时，WordPiece 子词切分采用从左到右的最大匹配方法，也就是一种贪心方法，非常简单。实际上 BPE 的子词切分也可以采用这种方法，效果与自底向上合并的方法上应该没有太大区别。

6.2.4 Unigram

Google 还提出了另外一种词表构造方法，叫做 Unigram 方法[299]。Unigram 和 WordPiece 一样，也采用了语言模型来决定词表中是否收录某个子词。WordPiece 原始论文 Japanese and Korean Voice Search 并没有明确说明所使用的语言模型，而 Unigram 方法使用的是一元语法语言模型，这也是这种方法取名为 Unigram 的原因。不过与 BPE 和 WordPiece 不同的是，Unigram 构造词表的方式不是先构造字符级别词表，再通过合并操作逐渐往词表中增加新的子词；相反，Unigram 先构造一个大的子词词表（比如收录语料库中所有词），然后再逐渐删除词表中的子词（把罕见的子词切分成两个更常见的子词），直到词表长度达到预定值。Unigram 词表构造方法在一个开源的工具 SentencePiece 中提供了具体实现，但运行效率比较低，实际上使用这种方法的人很少。

推理时，Unigram 的子词切分也采用从左到右的最大匹配方法。

6.3 SentencePiece

仔细研究上面介绍的子词切分方法会看到，我们在做子词切分之前，还需要先做一次词语切分（Word Tokenization），以得到一个初始词典（也就是前面 BPE 词表构造方法中的 Dictionary）。但词语切分也并不是一件简单的事情，比如：

◆在英语中，标点符号和字母、数字之间通常是没有空格的，尤其是英语的句号（.）和数词中的小数点、缩写词中的点号完全相同，需要专门的算法来排除歧义；

◆在中日韩等语言中，字符（主要是汉字）之间是没有空格的，需要专门引入外部的词语切分工具进行词语切分；

◆每种语言都有各自的词语切分问题，词语切分无法做到语言无关。

SentencePiece[300]就是为了解决这一问题提出的子词切分方法。其基本思想是，对于输入的文本无需先做词语切分，直接在输入文本串的基础上构造子词词表和进行子词切分。直观地理解，可以认为输入的每个句子（甚至段落）就是一个单词，然后再构造词表进行子词切分。在这个方法中，空格不作为词语之间的分隔符；也就是说，空格与其他任何字符同等对待。按照这种方式得到的子词词表，子词中出现空格，或者子词中同时出现字母、数字或者标点符号都是很常见的。其实验表明，采用 SentencePiece 的方法，无需先做词语切分，可以取得与先做词语切分方法可比的效果。

SentencePiece 也可以用于处理中文，无需事先做中文分词就可以直接使用。也有些人先对文本做中文分词，然后再调用 SentencePiece，也是完全可以的。先做词语切分，后面调用 SentencePiece 和直接调用 BPE 的效果相同。另外要特别注意推理时，也需要先用相同的中文分词工具做词语切分，然后再调用 SentencePiece 做子词切分，否则训练和推理子词切分过程不同，会导致效果很差。

SentencePiece 和 WordPiece 这两个名词容易造成混淆，让人以为 SentencePiece 是和 WordPiece 一样的一种子词切分方法。实际上这是两个不同层次的概念。WordPiece 和 BPE、Unigram 是类似的技术，都是用于构造子词词表和做子词切分的方法。SentencePiece 方法只是强调子词切分之前无需做词语切分，直接从句子开始构造子词词表和进行子词切分，但并没有规定子词词表构造和子词切分的算法。所以，即使采用 SentencePiece 方法，还是需要选择 WordPiece、BPE 或者 Unigram 来做子词词表构造和子词切分的。

另外要注意的是，SentencePiece 不仅是一种子词切分方法，也是一套开源的工具。这套工具提供了 BPE 和 Unigram 两个子词词表构造的工具。

6.4 子词正则化方法和 BPE-Dropout 方法

做过中文分词的人都知道，词语切分是有很多歧义的，子词切分也不例外。如果深入观察 BPE 或者 WordPiece 产生的子词切分结果，会发现很多不合理的切分。

BPE 切分方法还会导致一些子词训练不充分的问题。比如一个串 ABC 有可能切分成 A/BC 或者 AB/C，其中 A/BC 是合理切分、AB/C 是不合理切分。但由于语料库数据分布的问题，很可能导致语料库中 AB/C 出现的次数远高于 A/BC，这样 BC 虽然是一个高频子词，但在训练数据中出现的次数却非常少，从而可导致 BC 这个子词的表示训练得非常不充分。以表 5-1 为例，w+e 实际上是一个高频的子词序列，但我们学到的子词词表甚至没有学到这个子词。

Kudo 提出了一种子词正则化的方法来解决子词切分歧义的问题[301]。其思想还是引入一个语言模型（如 Unigram），对于多种切分结果，根据语言模型的概率随机取样其中一种切分结果，这样会使得切分结果比较合理。但这种做法的代价也很大，因为需要引入一个语言模型，会导致训练过程中的计算量大增。

论文 Language Models are Unsupervised Multitask Learners 提出了一种更加简单的 BPE-Dropout 技术，用于改善 BPE 方法的子词表示训练不充分的问题。其思想非常简单，在语言模型训练阶段做子词切分时，以一定的概率随机跳过一些子词合并操作，通过增加子词切分的随机性，就可以使一些原来训练不充分的低频子词出现概率大大增加，从而学到更好的子词表示。

另外要注意的是，无论是子词正则化还是 BPE-Dropout，这种在

切分过程中引入随机性的做法，通常都是只在模型训练阶段使用，以确保得到更好的模型；在语言模型应用（推理）阶段，为了保持一致性和提高效率，一般还是采用确定性的子词切分方法。

6.5 Byte-level BPE

6.3 节中提到了子词切分的方法，对于中日韩这样的大字符集语言来说，还是有问题的：字符太多，甚至可能超过预定的词表大小，导致一些字符无法收入子词词表，仍然会有 OOV 问题。即使字符集足够大，把所有字符都收入子词词表，会导致词表中收入大量罕见字，词表空间利用率大大降低；而在解码时，需要在词表的所有子词空间上计算 softmax，大量的罕见字出现概率极低，导致解码的时间效率大大降低。

论文 Training Multilingual Pre-Trained Language Model With Byte-Level Subwords 给出了一个 mBERT 子词词表（WordPiece 词表）中，每个子词在一个训练语料库中出现的频率分布图（见图 6-1）。从图中可以看出，频率分布在超过词表长度 60%以后快速下降，在超过 83%以后直接降到 0。说明 BERT 的词表使用非常低效。即使这样 mBERT 仍然存在 OOV 问题。我们在使用中发现，著名作家“章诒和”的“诒”如果写成繁体“詒”，mBERT 就无法识别。

Byte-level BPE（BBPE）为解决这一问题提供了有效的办法。BBPE 的基本思想是把所有字符表示成 UTF-8 格式，把 UTF-8 字符串当成一个字节组成的序列，然后以一个字节为最小的组成单位，构造 BPE 子词词表和进行子词切分。这样 BPE 最基本的组成单位最多只有 256 个，不存在大量罕见字符占用子词词表空间的问题。

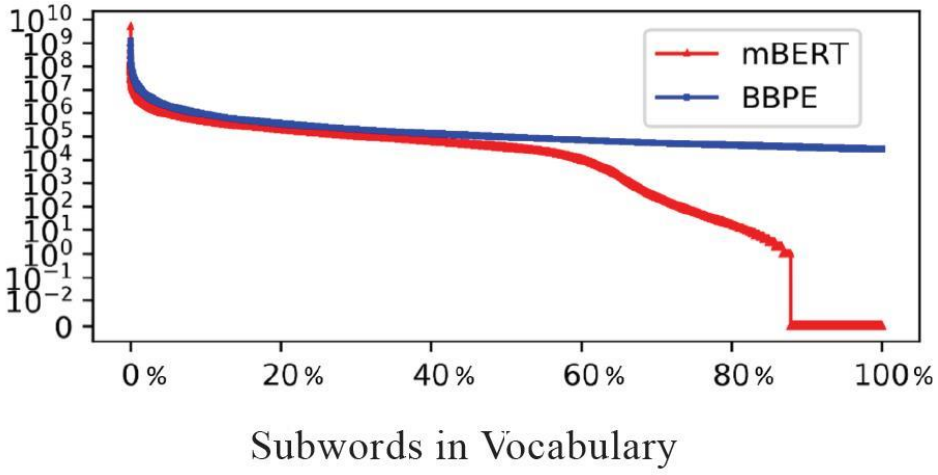


图 6-1 mBERT 词表和 BBPE 词表中的子词在训练数据中出现的概率分布

根据 UTF-8 的编码规则，一个字符可以表示成 1~4 个字节，长度不等。这样会带来一个问题，一些罕见字符会被切分成两个或者多个子词，理论上有可能导致生成时会出现非法字符（非法子词序列），但实际上这种情况几乎不会出现。因为在训练数据中几乎不会出现非法子词序列，这种概率分布是会被神经语言模型学习到的，模型生成时也会遵从这种概率分布，因此生成非法子词序列的可能性虽然理论上存在，但概率极低，实际上几乎不会发生。我们在使用 BBPE 训练 GPT 模型时，也从来没有遇到过生成非法字符的情况。

BBPE 方法首先在 GPT-2 模型中被采用，后来 RoBERTa 模型也采用了 BBPE 方法做子词切分，随后的 GPT-3 模型也延续了这种方法。论文 Neural Machine Translation with Byte-Level Subwords 和 Training Multilingual Pre-Trained Language Model with Byte-Level Subwords 分别对 BBPE 在神经机器翻译和预训练语言模型中的使用效果进行了深入分析。Huggingface 工具包也提供了 BBPE 模型的具体实现。华为诺亚方舟实验室在开源预训练语言模型代码库中也提供了 BBPE 的完整代码；另外考虑到已有的 BBPE 实现都没有对 SentencePiece 的支持，而 SentencePiece 使用的人仍很多，我们的代

码库中还提供了基于 BBPE 的 SentencePiece 的实现方案。

6.6 VOLT

在构建子词词表时有一个问题，就是子词词表规模要选择多大比较合适？理论上，词表越大包含的信息越多，模型效果当然更好；但模型越大，占用的空间就越多，推理时间也会相应增加，代价也会越来越高。那么在词表大小和系统性能之间是否存在一个合理的平衡点？论文 Vocabulary Learning Via Optimal Transport for Neural Machine Translation 提出了一种 VOLT 模型[302]，对这一问题给出了一个明确的答案（该论文获得了 ACL 2021 的最佳长论文奖）。

由于下游任务的性能衡量并没有统一方法，论文采用词表的信息熵来粗略代表下游任务的性能，词表越大，词表对应的信息熵就越低，相应的下游任务的效果就好越好；反之亦然。这样问题就转化成如何找到一个最优的子词词表，可以在词表的信息熵和词表的大小之间寻找一个合理的平衡点。

为了寻找这个最优平衡点，该论文引入了边际收益的概念，即将词表的增大理解为投入，词表信息熵的降低理解为收益；定义了词表边际收益（MUV）这个衡量指标——每增加固定的词表规模带来的信息熵降低的幅度。在词表很小时，加大词表带来的边际收益很高，随着词表越来越大，这种边际收益会越来越低。

为了获得一个 MUV 最大的词表，该论文把词表搜索问题转成了一种最优运输问题，此问题可以在多项式时间内用动态规划方法求解。在 52 个方向的机器翻译实验表明，采用此方法可以将词表规模降低到基线系统的 30%，但机器翻译的性能总体持平甚至略有提高。

6.7 形态切分

形态切分是通过分离词素解决因形态过于丰富而造成文本数据

稀疏的重要步骤，相比传统的子词切分，形态切分利用标注数据对模型进行有监督训练，拥有更高的切分准确率。

6.7.1 维吾尔语形态切分概述

以维吾尔语形态切分为例，维吾尔语是一种典型具有粘着语特性的拼音文字，主要使用人口分布于中国新疆、中亚与世界各地，单词由词根与词缀组合而成，词根蕴含单词的基本意义，词缀蕴含单词的语法含义[297]。构词时，在符合语法条件的情况下，单个词根可以和多个词缀自由结合构成新词，如词根 Beijing 可接不同词缀构成新词 Beijingda、Beijinggha、Beijingghicha、Beijingdin、Beijingni 等。据统计，词根ئاختۇر（搜索）与不同词缀相结合能够衍生出 3000 多个维吾尔语单词（常用汉语或英文单词才 3000-5000 多个）,而维吾尔语总计有 40000 多词根（不包括外来词）。因此，同等字符级的文本，维吾尔语文本词表数量要远远大于其他语言，在现有的 17 万条句子级维吾尔语语料库中统计其词表达 16 万。巨大的词表造成了严重的数据稀疏问题进而大量的未登录词问题也伴随而来，影响这类语言机器翻译等下游任务的研究。维吾尔语下游任务中，需要使用形态切分分离单词的词根与词缀，例如将 Beijingda->Beijing+da，Beijinggha->Beijing+gha，这样将有效减少词表数量，缓解数据稀疏问题。

6.7.2 维吾尔语形态切分前沿综述

维吾尔语信息处理任务已经有几十年的发展历程，研究方法大致可以分为三类：基于规则的方法；基于统计的方法；基于深度学习的方法。

1. 基于规则的方法

2002 年古井拉·阿东别克等人[303]提出使用词典查询、最大匹配完成音变还原、音节切分与词干提取任务，此方法考虑了语音规律与结构规则，同时也存在一定的歧义切分问题；2006 年陈鹏等人[304]提出基于语料库的提取办法，采用全切分、双向匹配算法与词典查询的

方式完成音变还原与词干提取任务，此方法尽可能的考虑了所有的切分形式同时对词干表和附加成分表规模要求较高；2008 年艾山·吾买尔等人[305]提出融合 FSM 和词典查询的算法，该方法使用 FSM 与词典查询完成音变还原和词干提取任务，有效减少了切分次数，速度快，缓解了未登录词问题，央音产生形态还原问题还有待解决。2009 年 ORHUN M[306]提出基于规则的名词分析方法，使用 Two-Level、有限状态工具与词典查询，基于规则的名词方法有效处理了名词形态分析，但对于词典的规模要求较高，歧义现象较难解决，易出现过度切分现象。

2. 基于统计的方法

2009 年 Batuer Aisha[307]针对传统规则方法过度切分现象提出了基于统计的形态切分方法，包括两步切分的统计方法和字母标记方法，主要使用 MEM、CRF 等模型。基于统计的方法充分利用了上下文信息，但是处理数据稀疏问题欠佳且特征设计困难。2010 年艾山·吾买尔[308]提出央音元音识别模型，使用信道噪声模型和 FSM 完成音变还原和词干提取任务，在处理弱化词方面效果良好，处理外来词结构能力欠佳。2012 年麦热哈巴·艾力等人[309]提出了音变现象自动还原模型，使用词内字母对齐算法与 MEM 解决了维吾尔语词素之间的音变还原问题，此方法不依赖于规则，特征模板存在局限性模型适应能力弱。2014 年张海波[310]提出联合形态分析方法，利用感知机和词内字母对齐算法完场形态切分任务，解决维吾尔语词干与词缀之间音变还原问题和形态切分之间错误传播问题，其处理联合标签导致系统速度缓慢。2015 年米尔阿迪力江·麦麦提等人[311]提出了 Morfessor,有效处理了维吾尔语词干提取与音变还原的特例和歧义问题。2016 年 TURSUN E[312]提出了半监督标签转换马尔可夫模型，使用马尔可夫、词典查询与标签转换的方法实现词干提取与音变还原解决了标签歧义问题并用较少词素分词，对人力需求大。

3. 基于深度学习的方法

2017 年哈里旦木·阿布都克里木[313]提出双向门限递归单元神经网络，将深度学习应用到维吾尔语形态切分领域，通过上下文信息消除切分歧义，并开源了实验代码与语料库。2018 年吐尔洪·吾司曼[314]提出字符级形态协同分析方法，利用序列标注方式完成多种形态分析任务，其忽略了单词间上下文的关系。2019 年 Yaofei Yang 提出 Point-erNetwork 形态切分方法，不同于“bmes”标注工作，该方法将较少的独立且包含全面信息的标签（即“b”和“s”）用于形态切分，有着较好的稳健性。2020 年 ABULIMITI A[315]提出关联语言模型，使用相关联的高资源语言改进形态切分模型，大规模维吾尔语数据的条件下，提升模型性能效果有限。

6.8 多语种新词发现方法

新词发现是自然语言处理领域的基础任务之一，通过对已有语料进行挖掘，从中识别出新词。本节主要讲述新词发现的定义以及多语种新词发现的前沿综述。

6.8.1 新词发现概述

语言随着社会的发展而发展，在词汇中的一大体现就是新词语的出现。自古以来，汉语词汇带有特定时代的烙印，必然会从一个侧面反映出社会政治、经济、文化以及人们价值观念、生活方式的变迁等。例如，在 20 世纪七八十年代，“工分”、“粮票”、“布票”等是人们耳熟能详的名词。随着改革开放的深入，新的词语不断出现在生活当中，如“科学发展”、“以人为本”、“政务公开”、“笔记本电脑”、“虚拟现实”等，这些词语真实地反映了社会和经济的飞速发展以及对外交流的日渐频繁。

目前，在自然语言处理领域中出现了新词和未登录词两种概念。通常未登录词被定义为未在词典中出现的词。新词虽然也是未在词典

中出现的词，属于未登录词，但它和未登录词还是不同的。新词一般是一个加入了时间的动态概念，未登录词是相对于词典的概念。这里从两个方面来把握新词的定义。

（1)从词典参照的角度来说,新词语是指通过各种途径产生的、具有基本词汇所没有的新形式，新意义或新用法的词语。

（2）从时间参照角度来说，新词语是出现在某一时间段内或来自某一时间点以后首次出现的具有新词形、新词义或者新用法的词汇。

新词作为一类网络流行语，其具有以下几种典型的特征。

（1）新颖性。新词最主要的特点在于“新”，符合当下时代发展的趋势。无论是从已有词语中演变而来，还是用户创造性地提出，这些新词都具有了新的含义，表达了新的思想。

（2）周期性。新词的产生一般依托于时下热点话题讨论，通常情况下一些新词随着事件热度的淡去便也渐渐消亡,但也有一些新词被保留下来。因此，不同的新词具有不同的存在周期。

（3）传播速度快。新词基于网络平台产生，并能借助网络平台迅速传播。同时，新词含义一般简单直接，易于理解。因此，可以在很短时间内被人们接受并在各个场合使用。

（4）不规则性。新词的构成比较自由随意，没有固定的格式要求，也不完全符合成词的规则，出现了一些新颖的构词方式，同时在长度和构词符号等方面也没有限制。

目前，新词识别主要研究方法：基于规则，基于统计和规则与统计相融合的 3 种方法。基于规则的方法利用构词学原理，配合语义信息或词性信息来构造模板，然后通过匹配来发现新词；基于统计的方法是通过对语料中的词条组成或特征信息进行统计来识别新词。基于规则方法的优点是准确率高,针对性强,但手工编写和维护规则困难,且规则一般是领域相关的，所以适应性和移植性比较差；基于统计方法的优点是灵活，适应能力强，可移植性好，但需要大规模语料进行

模型训练，由于使用的语言知识较少，一般都存在数据稀疏和准确率低的问题。目前大部分研究者使用规则和统计相结合的方法，以期发挥各自的优势。

6.8.2 多语种新词发现前沿综述

已有的新词发现算法大致有以下两种：一种是基于构词法的算法，也叫做基于规则的算法。这种方式基于语言特征构建的规则库，规则的构建过程往往比较复杂，并且模型的迁移能力比较差。

另一种是基于统计的新词发现算法，目前主要分为以下两类：（1）基于对语料库的频繁模式的发现。Huang 提出了一种使用邻接熵和互信息作为特征进行新词发现的算法[316]。此类算法需要涉及频繁项的迭代发现以及上下文信息的获取，时间复杂度和空间复杂度较高，不适合大规模语料的处理。（2）使用标注模型进行新词发现。Peng 使用 CRF 模型计算汉语片段的置信度在分词的同时提取新词[317]。这一类的算法基于一个词上下文的局部特征，准确率不高。2017 年，Zhang 将上述两种方式结合[318]，使用 CRF 模型进行候选词提取，使用二元语法模型重新扫描语料，提取候选词集左右熵、互信息等特征。该方法能够有效避免传统算法中对全局状态的依赖，实现了对大规模语料的快速新词发现。

在新词发现的特征选择方面，Luo 比较了九种常见的词内部特征计算方法[319]，实验表明使用互信息的效果最好。Huang 提出了一种基于模式的框架[320]，将这些统计特征整合在一起来检测新单词。

近年来，预训练语言模型有很大突破，这些模型已被证明能有效地解决各种各样的任务。对于新词发现这一问题，2019 年，McCrae 提出了一个基于预训练语言模型的“形容词+名词”新词短语识别分类器[321]，实验结果表明深度学习模型结合频率特征的效果最好，但这种方式没有考虑到短语的上下文信息。

在新词发现的噪声词过滤方面，2017 年，Liang 从新词外部环境

稳定性的角度，定义了 overlapping score 来过滤噪声词[322]。2018 年，Zhang 利用词向量构建弱成词词串集合来过滤成词能力较弱的候选词[323]，其性能超过了 overlapping score 的效果，并表明使用包含词内位置信息的字向量的过滤效果最优。2019 年，Qian 提出了 WEBM 模型[324]，基于词向量计算词碎片的余弦相似度，设置相似度阈值对噪声词进行过滤，实验结果表明 WEBM 在从大量中文语料库种检测新词方面具有很大的优势。2022 年，Zhang 等人在 WEBM 的基础上提出了 MWEC 模型[325]，引入外部知识库训练多语义词向量，并应用到候选词集剪枝中，解决了中文的一词多义问题。

目前对于新词发现的研究主要集中在现代语言语料，在古语语料中的研究鲜有涉及。2017 年，Xie 提出了 AP-LSTM 算法[326]，是专门针对古汉语语料的有监督新词发现算法。2019 年，Liu 提出了古汉语的新词发现算法 AP-LSTM-CRF[327]，利用数据挖掘的关联规则算法和深度学习的方法有效地挖掘古汉语语料中的新词，并在宋词和宋史数据集上验证了模型的有效性。

在实际应用中，特别是对于多语种而言，获取大量的标注语料十分困难，因此大多数学者致力于探索无监督的挖掘方法以实现新词探索。

2008 年，Humbley 实现了 NEOROM：一个针对拉丁语言的新词检测系统[328]。2017 年，Cartier 实现了一个能够自动识别 7 种不同类型语言（中文、捷克语、法语、希腊语、俄语、波兰语、葡萄牙语和斯拉夫语）新词的系统[329]，通过报纸语料库来跟踪新词的生命周期。对于一些东亚语言，如汉语、日语、泰语等，词与词之间没有明确的边界。2015 年，Uchiumi 等人提出了一个非参数贝叶斯模型[330]，直接从字符串构建类 n-gram 语言模型，同时集成字符和单词级别的信息，在日文、中文和泰文的标准数据集上的实验结果显示，该算法的精度优于以往的结果。2014 年，Falk 基于统计的方法对法语语料进

行新词发现[331]，将任务转化为有监督的分类问题，并讨论了三组特征的影响：形式相关特征、形态-词汇特征和主题特征。2018 年，Klosa 提出了一种半自动的德语新词检测方法[332]，并探讨了对于专业词典编纂的影响。

6.9 小结

本章主要综述了基于神经网络的 NLP 方法中采用的子词切分方法。首先阐述子词切分问题的来由，封闭词表假设所带来的集外词（OOV）问题，以及为什么要采用子词切分来解决这一问题。然后介绍构造子词词表和子词切分的常用方法，包括 BPE、WordPiece 和 Unigram。然后提出子词切分技术所面临的一些相关问题及其解决办法，包括：子词切分之前需要先进行词语切分的问题，以及避免该问题可以采用的 SentencePiece 方法；切分歧义问题和子词表示训练不充分问题，以及解决这一问题的子词规范化方法和 BPE-dropout 方法；大字符集问题，以及解决这一问题的 BBPE 方法；词表最优化问题，以及解决这一问题的 VOLT 方法，最后阐述了形态切分与多语种新词发现。

子词切分是神经 NLP 方法的最基本技术之一，全面了解这一技术有利于更好地理解基于神经网络的 NLP 模型，并设计更好的 NLP 系统。

本章编写人员：

刘群、杨子妍、严若豪

第 7 章 多语种机器翻译

7.1 多语种机器翻译概述

多语种机器翻译又称“多语言机器翻译（multilingual machine translation, MMT）”。自 2014 年以来，端到端及基于注意力机制的神经网络机器翻译（neural machine translation, NMT）模型日趋成熟，其翻译性能相比于传统的统计机器翻译模型有了很大的提升。神经网络机器翻译这一范式问世不久，人们就发现这一框架可以自然地适用于多语言场景下，因此就有很多工作都在研究多语言神经翻译系统。拥有高质量、大规模的两种或多种语言之间的平行语料是当今神经网络机器翻译模型获得良好性能的前提。然而，除了英文、德文、中文等少量几种资源丰富语言外，世界大多数语言都无法找到大规模的双语平行语料以满足神经机器翻译模型的需求。当今世界有 7000 多种语言，机器翻译系统研究与开发应用大多集中于诸如英语或其他具有大规模文本语料库的几种语言中（约 20 种左右），世界其他大多数语言都急需相应的语言处理工具和语料资源以满足当前深度学习模型的计算需求，这些语言被称为低资源语言（Low-Resource Languages）。如何解决低资源语言的机器翻译问题成为当今机器翻译研究的热点领域。

若一个神经网络机器翻译系统能处理多于一个语言对之间的翻译，就可以被称作多语言神经网络机器翻译系统（multilingual neural machine translation, MNMT）。MNMT 的终极目标是一个能够有效利用可用的语言资源，翻译尽可能多的语言的机器翻译系统。在这个系统中，由于知识迁移的作用，低资源语言的翻译可以从其他语言对中获取额外知识，导致翻译效果得到很大提升，且 MNMT 系统的泛化性会提高，应用更加广泛。根据其应用场景，MNMT 可以被分为三种：（1）多路翻译。系统使用统一的模型，利用多语语言对中一个语

言对的语料，训练得到的翻译系统可以做到一对多翻译、多对一翻译或多对多翻译。(2) 低资源翻译。多语言系统为低资源语言翻译提供了两种解决思路：利用迁移学习增强现有双语平行语料和利用枢轴语言融合语言模型。(3) 多源翻译。已被翻译成一种或者多种语言的文档在将来可能要被翻译成更多种语言。这种情况下，已经产生的目标语言内容可以为将来的翻译文本内容消歧从而提升翻译质量提供训练数据。

7.2 多路翻译

MNMT 的初衷是能支持多个语言对的互译，具有这种能力的模型称为多路神经网络翻译模型。以下工作分别从参数共享、训练方法和语言多样性等角度提出了各自的解决方案

7.2.1 参数共享

早期的多语言翻译模型是基于循环神经网络（Recruit Neural Network, RNN）的，例如在 Firat 等[333]的模型中，不同语言都使用独立的词嵌入、编码器和解码器，只共享注意力机制。此外，模型中加入了两个共享构件：一个对所有编码器都适用，以编码器的最终状态为输入，来初始化解码器初始状态；另一个仿射层对所有解码器都适用，然后在计算 softmax 之前将解码器最终状态进行投影。这种 RNN 模型虽然比较灵活，但是参数量大，而且所有语言对翻译都需要通过共享的注意力机制，容易造成表示瓶颈。

Google 在 2017 年提出了一种改进的模型 Google MNMT（GMNMT），所有语言共享词嵌入、编码器、解码器和注意力机制，但在每个句对前面加一个语言标签（language tag），以指明应该翻译成什么语言，帮助编码器正确生成对应目标语言的句子。若系统中涉及的语言有较近的亲属关系，模型效果会尤其好，因为它们有更高的词汇和语法相似性；再如果使用同一种书写系统，效果会更好[334]，

但在不同亲属关系或者不同书写系统的语言对的翻译中效果不佳。Ha 等人[335]提出了一种类似的多语言翻译系统，不过该工作为每个语言保留一个独立的词表，因此可能更适用于语言系属关系不强的情况。这种完全参数共享的方法简单直观，参数量最少，且与双语系统相比不占下风，有时评测指标 BLEU 值还会有所提升。在这个范式下，诞生了很多工作：有的工作通过处理语料，使得不同书写系统的语言能够更好表示（例如使用转写（script conversion）、音译（transliteration）、使用基于字符的模型等）；有的工作通过数据选择策略和语料平衡策略提升低资源语言的重要性；有的工作通过加深 Transformer 的层数提高模型的表示能力。总体来说，这种方法能通过知识迁移有效地提升低资源语对的翻译效果，但是整体上，将低资源语言翻译成资源丰富语言的效果要好过把资源丰富语言当做原文情况下的翻译效果。此外，在海量语料和大量语言的高强度机器翻译需求之下，完全参数共享的模型也存在表示瓶颈。Arivazhagan 等人[336]对该现象进行了更细致的论述，而 Kudugunta 等人[337]对多语言模型的工作原理进行了可视化，有助于更直观的了解。

作为上述两种方法的折中，一些工作试图使得不同网络层的参数共享程度可控，其背后的动机是多语言系统接受的语言通常存在差异性，以及希望网络有一定的灵活性和简易性。由于生成的工作主要由解码器完成，因此让解码器独立是比较重要的，相比而言，编码器的工作比较简单，所以大部分工作会选择多个语言共享解码器，达成更好的参数利用率。但在这种情况下，解码器和注意力机制要尽可能鲁棒。Blackwood 等人[338]指出目标语言特定的注意力机制比其他注意力共享方案效果要好，因此设计一个强壮的解码器尤为重要。Sachan 等人[339]则指出对于基于自注意力的模型，共享解码器自注意力和编-解码器注意力参数对不相似的语言效果更好，因为解码器学到的目标语言表示能更好地和源语言表示对齐。Bapna 等人[340]是在完全参数

共享的模型上作了扩展插入特定于语言对的适配器层（adapter layer）。宿主模型训练好以后，将额外的适配器层插入到模型中，对特定的语言对微调这一层。Zaremoodi 等人[341]则认为不应该在训练之前固定设置哪些参数应该被共享，因为共享的参数可能有助于提升一些语言对但是不利于另一些语言对，他们提出了一种路由网络（Routing Network），动态控制哪些参数应该共享的算法。Platanios 等人[342]则是从训练数据中学习参数的共享程度：其定义了基本参数和语言嵌入。对于给定的语言对，基本参数会通过线性投影变换为这两个语言各自的参数。相比较于语言标签，语言嵌入也能更直接影响模型参数。

设计正确的共享策略可以平衡模型的大小、翻译准确性、简易性和灵活性，因此非常重要。但并未有太多工作研究这些模型的表示瓶颈。另有一些工作使用强化学习或遗传算法来研究神经网络架构搜索（Network Architecture Search, NAS），并通过这种方法自动调整网络大小。此外，条件计算（conditional computation）可能是多语言翻译可用的一种技术。

7.2.2 训练方法

MNMT 模型的训练是一个关键问题，需要比较精妙的方法。所有方法的核心都是对所有语言对最小化负对数似然。主流的训练方法有两种：单阶段训练（或称并行训练、联合训练）和多阶段训练（或称顺序训练）。根据具体用例，多阶段训练可以用来做模型压缩或微调，或加入新数据/语言（增量训练）。

7.2.2.1 单阶段翻译

最简单的情况，把所有语料拼在一起然后送入模型。如果模型有多个编/解码器，每个 batch 包含的所有数据需要来自同一个语言对；如果模型参数完全共享，则不需要受此限制。单阶段翻译的最大问题是数据不平衡问题，传统做法是过采样低资源数据，但是近几年 Arivazhagan 等人[336]提出的基于温度系数的采样（temperature-based

sampling）方法开始流行。

7.2.2.2 知识蒸馏

原始的知识蒸馏方法是训练一个极深的大模型，然后原始数据和该模型预测得到的 softmax 组成新的数据，训练浅层小模型。这种方法后来得到改进，成为序列级知识蒸馏，就是用大模型翻译一遍训练集，然后把源句和翻译结果组成新的数据集训练小模型。Tan 等人[343]为所有语言对训练双语模型，然后把它们作为教师模型来为所有语言对训练一个学生模型。训练学生模型时，使用翻译损失和蒸馏损失两者的插值，其中后者捕捉学生模型输出的概率分布和教师模型输出的概率分布之间的距离。只有当教师模型在验证集上的准确率好过学生模型时，蒸馏损失才起作用。这种方法比联合训练效果好，但是代价太大，因为每个语向都需要一个单独的模型。

7.2.2.3 增量训练

这些方法主要是想减小引入新语言或新数据时的代价，至少要避免重新训练。新语言引入以后最直接的问题就是会修改词表，因此很多工作都聚焦于此。最简单的方法是对非拉丁字母语言使用拉丁转写，这样可以无缝继续训练或者微调模型。Bapna 等人[340]在预训练 MNMT 模型上使用小的前馈神经网络（称为适配器）。对预训练模型中的每个语言对，该工作在每一层都加入适配器，然后在该语对数据上微调这些适配器——但是这种方法只能解决新数据带来的问题，对新语言无能为力。此外，增量训练不可避免会带来灾难性遗忘（Catastrophic Forgetting），这还需要进一步研究。

对上述所有方法的一个主要的批评是它们把 MNMT 模型都看作是一种普通的 NMT 模型，很多研究人员都是平等对待所有语言对（除了在采样方式上有所不同），但是很少有人研究 NMT 是否实际上更擅长处理某些语言对，或者更不擅长处理哪一些。有工作研究如何对资源丰富度不同的任务缩放学习率或梯度。另外，现在的训练过程中

使用的开发集都是多语言的，因此选出的模型可能对某个个别的语言对并非最优——甚至大部分工作的模型都是次优的，即便它们已经对低资源任务有了相当大的提升。

7.2.3 处理语言多样性

MNMT 的一大核心任务是对齐不同语言的单词/句子表示，使得亲属关系较远的语言也能被连接起来，使得模型能够处理尽量多的语言。词表对 MNMT 模型很重要，但是目前并未得到太多重视。对共享词表的 MNMT 模型，很自然的解决方案是对每个语言采样等量单词来平衡表示能力，但是考虑到数据的不平衡现象，这种做法仍有改进空间。Aharoni 等人[344]提出了基于温度系数的采样方法，是一种当前被广泛使用的策略。以下主要讨论多语言的表示：

7.2.3.1 多语言表示的本质

已有工作对多语言模型的嵌入作了可视化，并认为编码器可以对不同语言的相似句子学到相似表示，但是这些可视化工作通常是在 2-3 维空间中做出，因此存在一定局限性。Kudugunta 等人[337]使用 SVCCA 对表示结果作了更系统的分析，认为尽管编码器能对不同语言的相似句子学出相似表示，但是这些表示可以根据语言相似度聚成更细粒度的若干簇。这也解释了为什么对相近语言迁移学习效果更好。编码器和解码器之间的界限不明显，源语言表示和目标语言表示互相依赖，不同层学到的表示相似性不同。编码器端，越高层学到的表示越不变，而解码器端相反，越高层学到的表示差异越大。这个结论符合直觉，因为解码器需要敏感于生成什么语言，也就需要在语言不变的表示和语言感知的表示间寻求平衡。

7.2.3.2 编码器表示

影响编码器表示，使其依赖于具体语言的原因有若干点，其中一点是对于同一个目标句，不同源语言对应的句子其词元数可能不同，因此解码器的注意力机制看到的编码器表示数量就会根据语言不同

而不同。一种方法是如 Lu 等人[345]、V á zquez 等人[346]建立一个注意力桥接网络，生成固定数量的上下文表示，这样解码器的任务被简化，可以更善于语言生成。Murthy 等人[347]则指出编码器产生的句子表示依赖于词序，因此是语言相关的。其提出的对应方法是重排序输入句的词元，进而提升迁移学习的效果。

7.2.3.3 解码器表示

对于一对多任务，解码器的设计更有挑战，因为它更要兼顾语言不变性和语言相关性。如果只是把平行语料拼在一起训练模型，那么由于语言不变性的作用和词表泄露，模型很容易解码生成混合语言。最有效的对策是使用语言标签。

Wang 等人[348]探索了三种支持多目标语言的方法：（1）在目标句句首加目标语言标签；（2）设计目标语言相关的位置编码；（3）将解码器每一层的隐藏单元划分成共享的或语言相关的。每种方法与基线模型相比均有提升，综合利用最好。Hokamp 等人[349]则认为为每个语言使用不同解码器和注意力机制能得到更好的结果，因此最好的设计实践可能是共享编码器而分离解码器。即便是使用共享解码器，使用任务相关的嵌入也比使用语言标签好。另外一些研究则是将不同语言按照语系/语族划分，不同语系/语族的语言独立使用解码器。本文认为一种折中方案是不同语系/语族的语言独立使用解码器，语系/语族内使用标签提示模型输出何种语言。但是较新的预训练工作 MASS、mBERT、XLM 等已经证明语言标签足够帮助大模型确定应该输出何种语言。已有工作尚未对语言之间的负面影响（干扰）进行更深入的研究。

7.3 低资源翻译

数据稀缺是低资源语言机器翻译面临的主要问题，解决这一问题有两个思路：一是充分利用既有的双语平行训练语料，这个思路都是

对现有数据的增强，我们称之为“数据增强（Data Argumentation）”。二是融合单语训练语料的语言模型和翻译模型，这个思路是在现有深度学习算法和神经网络模型基础上的融合，我们称之为“模型融合（Model Stacking）”。

7.3.1 增强现有双语平行语料

反向翻译[350]（又称“回译”，Back Translation）是目前机器翻译任务中最常见的一种数据增强方法，其主要思想是：利用目标语言-源语言翻译模型（反向翻译模型）生成伪双语句对，用于训练源语言-目标语言翻译模型（正向翻译模型）。反向翻译方法只需要一个反向翻译模型，就可以利用机器翻译产生的数据增加训练语料的数量，因此得到了广泛地应用。对于低资源语言机器翻译，通过将目标语言句子复制到源语言端构造出伪训练语料能够提升机器翻译的性能。即使构造的伪训练语料是不准确的，其目标语言端仍然是真实语料，这样就既保留了两种语言之间的互译信息，又存在一定的噪声。神经网络机器翻译模型在伪双语句对上进行训练，可以学习到如何处理带有噪声的输入，从而提高了模型的健壮性。由于反向翻译模型的训练只依赖于有限的双语语料，因此生成的源语言端伪语料的质量难以保证，为此可以采用迭代式反向翻译（Iterative Back Translation）的方法，不断通过反向翻译的方式提升正向和反向翻译模型的性能。如果只有源语言的单语语料，也可以通过一个双语语料训练的正向翻译模型获得相应的目标语言翻译，从而构造出正向翻译模型的伪训练语料。然而，由于生成的译文质量很难保证，构造的伪训练语料对译文的流畅性并没有太大帮助，其主要作用是提升编码器的特征提取性能。

词替换[351][352]（Word Replacement）是将双语语料中的部分词替换为词表中的其他词。通过替换词，在保证句子语义或者语法正确的前提下，将替换以后的句对添加到训练语料中去，可以增加训练语料的多样性。可以替换源语言中的词，也可以替换目标语言中的词；可

以替换常用词，也可以替换稀有词；可以“刻意”替换，也可以随机替换；可以替换掉一个词，也可以丢弃这个词或者用掩码屏蔽该词；可以用词表中的其他词替换，也可以用本句中的其他词替换。词替换方法的本质是对原始双语训练语料进行修改，得到加了噪声以后的伪双语训练语料，以上词替换方式都是对原始语料进行加噪处理。在神经网络机器翻译中，通过加噪进行数据增强的常用方法是：在保证句子整体语义不变的情况下，对原始的双语语料适当加入一些噪声，从而生成伪双语语料来增加原始训练语料的规模。

相比词替换方法，转述（Paraphrasing）就不只是对句子做轻微的修改，而是考虑到了自然语言表达的多样性。即通过对原始句子的进行改写，使用不同的句式来表达同样的意思。通过转述的方法对原始的语料进行改写，可以使训练语料覆盖更多的语言现象，在神经网络机器翻译任务中得到了广泛应用。同时，由于每个句子可以对应多个不同的译文，转述方法可以避免模型过拟合，从而提高模型的泛化能力。

7.3.2 融合单语语言模型

通常情况下，在机器翻译系统中单语语料会被用于训练语言模型，训练好的语言模型既可以用于源语言端，也可以用于目标语言端。在源语言端，语言模型可以用于句子编码和生成句子的表示；在目标语言端，语言模型会帮助系统选择更流畅的译文。在低资源机器翻译中，语言模型融合可以在一定程度上弥补双语训练数据稀缺的缺陷。

因为神经网络机器翻译模型的解码器在本质上也是一个语言模型，用于描述生成译文词串的规律，那么将解码器与目标语言端的语言模型融合就成为一种最直接的使用单语数据的方法。常用的融合方法分为浅层融合和深度融合，前者独立训练翻译模型和语言模型，在生成每个词时对两个模型的预测概率进行加权求和得到最终的预测概率；而后者则联合翻译模型和语言模型进行训练，从而在解码过程

中动态地计算语言模型的权重，计算预测概率。

同样，神经网络机器翻译模型的编码器也是一个语言模型，是源语言的语言模型，但编码器并不直接输出源语言句子的生成概率，因此可以使用更大规模的单语语料对编码器进行训练。可以直接使用一个预先训练好的编码器，与机器翻译的解码器配合完成翻译任务。这种方法叫预训练[353]（Pre-training），即把句子的表示学习任务从翻译任务中分离，从而利用额外的更大规模的单语语料进行学习，得到神经网络机器翻译模型中的部分模型的参数初始值。然后，在双语数据上进行调参得到最终的翻译模型。对每个独立单词进行的表示学习结果称为固定词嵌入（Word Embedding），但在不同上下文中，同一个单词经常表示不同的意思。模型需要通过上下文理解每个词在当前语境下的具体含义，因此就需要上下文词嵌入。和固定的词嵌入相比，上下文词嵌入包含了当前语境中的语义信息，丰富了模型的输入表示，降低了训练难度。然而，为方便提取整个句子的表示，模型仍有大量的参数需要学习。因此，大量的预训练模型被提出。目前，应用最为广泛的有生成式预训练[354]（Generative Pre-training, GPT）和来自 Transformer 的双向编码器表示[355]（Bidirectional Encoder Representations from Transformers, BERT）。预训练模型在资源充分的语言翻译中效果并没有明显提升，但在低资源语言机器翻译中效果却比较显著。这是因为预训练阶段的训练语料规模非常大，因此在下游任务的数据量较少的情况下帮助比较大，低资源语言机器翻译就刚好符合这个特征。机器翻译是一种典型的语言生成任务，不仅包含源语言表示学习的问题，而且包含序列到序列的映射、目标语言端序列生成的问题，这些知识无法通过源语言单语语料训练和学习得到，因此，需要使用单语语料对编码器-解码器结构进行预训练。

7.3.3 低资源翻译方法

尽管神经网络机器翻译模型在拥有大规模高质量平行语料的语

言对中取得了很好的性能，但是实验证明，神经网络机器翻译性能在资源匮乏的语言对中性能大幅降低，甚至不如传统的统计机器翻译模型。研究者们充分利用可以利用的语料资源，探索了多种在低资源场景下的神经网络机器翻译方法。本文将当前的低资源语言神经网络机器翻译方法进行归类，按照训练过程使用的语料类型，将低资源语言神经网络机器翻译方法分为：监督方法、半监督方法和无监督方法。

7.3.3.1 监督方法

低资源语言神经网络机器翻译方法中的监督方法，指的是在整个模型训练过程中，需要直接提供源语言和目标语言之间的双语平行语料。监督方法又可以进一步可以分为以下四种方法：反向翻译（Back Translation）、词替换（Word Replacement）、迁移学习（Transfer Learning）和元学习（Meta-Learning）方法。前两种方法侧重增加训练语料，在前文中已有提及，此处不再赘述。值得一提的是，反向翻译方法在国内外机器翻译评测比赛中已经被认为是提升机器翻译性能必不可少的步骤。

迁移学习方法指的是利用从已知任务中获得的知识来改进相关任务的性能，这种方法通常可以减少所需的训练数据量。对于神经机器翻译模型的迁移学习[356]，其主要思想为：首先，训练在资源丰富语言对中训练一个神经机器翻译模型（父模型）；然后，使用父模型的神经网络参数初始化和约束低资源语言对（子模型）的模型训练。

元学习算法是“快速适应新数据”的最有效算法。Gu 等人[357]首次将元学习算法运用在神经机器翻译任务中，其基本思想为：首先，在多对资源丰富语言平行语料中训练高性能的神经网络机器翻译模型；其次，构建一个所有语言的词汇表；最后，根据词汇表和模型参数进行低资源语言神经机器翻译模型的初始化。可以看到，元学习模型更像是迁移学习运用在神经机器翻译中的进一步深化表示，实验结果也表明，元学习神经机器翻译在低资源语言对中表现出很好的性能。

虽然监督方法在很多翻译任务中取得了很好的性能，但同样存在一些局限：首先，该方法的最终性能很大程度取决于利用现有平行语料训练的机器翻译的质量，不适用于零资源语言（无平行语料的翻译任务）；其次，该方法运用在低资源语言机器翻译任务中存在一些语言自适应问题，即语言本身的特点对最终模型性能也有一定的影响。

我们使用数据增强技术初步尝试了有监督方法的工作，不同于传统的反向翻译方法，提出了一种基于语义相关词替换的数据增强方法。我们的目标是在保持源语言和目标语言句子结构的前提下，丰富两种语言句子的语义信息以达到扩充语料的目的。具体地，首先，我们训练目标语言的词向量模型选取语义相关词；其次，我们根据选取的语义相关词结合语言模型寻找语义相近的替换词；然后，根据词对齐模型生成扩充的双语平行语料；最后，我们对生成的伪语料进行语法纠错。实验结果表明，所提的基于语义相关词替换的方法能达到甚至超过当前数据增强技术的基线。

7.3.3.2 半监督方法

低资源语言神经网络机器翻译方法中的半监督方法，指的是在整个模型训练过程中，不直接使用双语句对的平行句对，而是采用间接的方式利用双语语料库。半监督方法又可以进一步可以分为以下两种方法：枢轴方法（Pivot-Based）和双语语料挖掘方法（Parallel-Extraction）。

枢轴方法[358]指的是利用第三种语言作为枢轴语言（通常为英语），搭建枢轴语言与源语言和目标语言之间的机器翻译模型，进而构建源语言与目标语言之间的机器翻译模型。该方法的主要步骤为：首先，训练源语言与枢轴语言（记作 S-P）和枢轴语言与目标语言（记作 P-T）的神经机器翻译模型；然后，利用 S-P 和 P-T 机器翻译模型将源语言翻译到目标语言，形成源语言和目标语言之间的平行语料；最后，根据构建出的源语言与目标语言之间的双语语料进行机器翻译模型的

训练。基于枢轴的神经机器翻译方法在零资源语言对（两种语言之间没有直接的平行语料）取得出很好的性能，是零资源机器翻译的最主要方法。

双语语料挖掘[359]指从大规模的互联网数据中挖掘出双语平行语料进行机器翻译的一种方法。该方法主要步骤为：首先，从网络中提取大规模的语料（非平行对齐语料）；然后，将这些语料通过现有方法（知识库、多语言句子嵌入等）进行双语平行语料挖掘；最后，使用挖掘生成的平行语料构建神经网络机器翻译模型。双语语料挖掘是构建机器翻译所需平行语料的主要手段，是从无到有搭建零资源语言（尤其是濒临灭绝的语言）机器翻译的有力手段。

半监督方法在挖掘平行语料资源方面取得了很好的效果，但是该方法存在一定的局限性：间接使用的平行语料的质量得不到保障，存在明显的错误传播问题，即神经网络机器翻译模型的最终性能会随着错误语料的堆积而放大。

我们开展了基于枢轴语言的神经网络机器翻译研究工作，借鉴传统的枢轴机器翻译方法，结合对偶翻译模型和模型融合技术。与传统的多次翻译的枢轴翻译相比，我们的目标是尽量减轻多次翻译累积的错误。我们在爱沙尼亚语、拉脱维亚语、罗马尼亚语与汉语的翻译任务中进行了实验，实验表明，所提方法大大超过了传统的多次迭代翻译方法。

7.3.3.3 无监督方法

低资源语言神经网络机器翻译方法中的无监督方法，指的是在整个模型训练过程中，不使用双语句对的平行句对，仅利用源语言和目标语言中的单语数据搭建神经网络机器翻译模型的一种方法。

无监督神经机器翻译方法[360]是仅利用两种语言的单语数据进行神经网络机器翻译模型搭建的一种方法。该方法一般有如下三个步骤：1) 利用大规模的单语数据训练跨语言词嵌入（Cross-Lingual Word

Embedding)，根据跨语言词嵌入初始化一个源语言到目标语言的机器翻译模型；2) 利用两种语言大规模单语数据分别训练源语言和目标语言的语言模型，作为降噪自编码器（De-noising Auto-encoding)；3) 利用反向翻译将无监督机器翻译问题转化为监督机器翻译问题，并进行多次迭代。无监督神经机器翻译方法的提出，在机器翻译领域引起了轰动，它颠覆了传统的机器翻译训练必须依赖平行语料的限制，并在两种语言比较相近的语言对(例如，英语和德语)中取得了很好的性能。

无监督神经网络机器翻译在一定程度上颠覆了人们对机器翻译研究的认识，在零资源机器翻译任务中取得了很好的效果，但实验表明，该方法仅仅在语言相似的两种语言对中取得了很好的性能，在远距离语言对中的性能极差。

我们在最新提出的无监督神经网络机器翻译模型的基础上，对翻译模型进行了改进。具体地，首先，我们在大规模的维基百科语料中训练多语言句子嵌入模型挖掘平行语料；然后，我们从挖掘出来的平行语料中抽取双语词典作为监督信号指导跨语言词嵌入的生成；最后，我们将训练好的跨语言词嵌入对无监督神经网络机器翻译模型进行初始化。我们分别在阿拉伯语、俄语、葡萄牙语、印度语与汉语的翻译任务中进行实验，实验结果表明，我们提出的方法可以提升无监督跨语言词嵌入的性能，但受限于这些语言与汉语之间的差异较大，最终在无监督神经网络机器翻译中仅有微弱的提升。

7.4 多源翻译

如果源句已经被翻译为多个语言，这些翻译结果可以一起用来改进到新目标语言的翻译质量，这种技术称为多源翻译（Multi-source Neural Machine Translation 以下简称 MNMT)。其底层原理是源语言的语言现象在多语言之间表达的形式不同，可以利用他们的互补信息。

7.4.1 多源翻译的发展契机

人们直觉上会认为获得同一个原文的多个不同目标语种的译文是不现实的，但实际上在欧盟和印度，这种情况是存在的。欧盟的官方语种数量超过 10 个，印度的官方语种超过了 22 个。欧洲议会的进行均是以多语言进行的，尤其是涉及到影响多个成员国的问题时。因此这些议会的内容会被人类翻译成多个不同语种已经成为了非常普遍的现象。基于这种现象，与其针对每个语种训练一个翻译器，不如针对这些语种的子集训练翻译器，然后利用这些子集语种的多源机器翻译系统进行翻译，这种多源翻译系统往往相对于单源的翻译系统能产生翻译质量更高的译文。这些多源机器翻译系统产生的译文可以进一步被译员进行以后编辑，然后输出成为多源机器翻译系统的训练语料库中。这样就可以产生类似于欧盟语料库和联合国语料库的多语平行语料库。

7.4.2 可获得多源数据

大部分研究假设同样句子会被翻译成不同语言，尽管这不一定是真的。但是，只要多源句子可用，就应该尽量用好。和多路神经网络机器翻译模型类似，多源神经网络机器翻译模型可以包含多个编码器，也可以只包含一个。Zoph 等人[356]表明，每个源语言使用独立的编码器和注意力网络构成的多源神经网络机器翻译模型比单源系统好。该系统中，每个编码器为不同的源语言句子生成不同的表示，解码器对每个源语言有独立的注意力机制，它们被拼接起来以后送入解码器，因此解码器的隐藏层会特别大，参数也很多——因此，可以考虑不用拼接，而是使用线性变化降维拼接得到的上下文向量，来避免解码器维数过大的问题。Firat 等人[361]并没有使用特别的多源模型，而是直接使用多语言模型。它也为不同的源语言生成上下文，但是并没有连接和投影，而是简单将它们相加作为解码器的输入——这种方法称为“早平均”；对应的，“晚平均”是解码器在每一时间步使用各个源语

言计算 softmax 然后求平均。将这两种方法结合，效果最好。

Dabre 等人[362]直接将多个源句拼在一起，送入标准神经网络机器翻译模型，效果与 Zoph[356]可比。有趣的是，这个模型可以自动判别不同语言句子之间的边界，简化训练过程。此文还显示最好使用语言相近的源语言，注意力机制更倾向于选择某些语言。如果两个语言亲属关系较远，则解码计算上下文时更容易被忽略掉。Garmash 等人[363]提出使用多个模型做模型组合的方法，学习一个组合函数来组合多个双语模型的 softmax 结果。这种方法需要的 N-语语料更小，但是训练组合函数可能比较费事。然而，Dabre 等人[362]指出普通的组合方法也可以得到可观的改进，这种方法相当于 Firat[361]的“晚平均”方法。

7.4.3 多源数据的缺失

针对多源平行语料中可能存在部分源句子缺失的问题，Nishimura 等人[364][365]先后提出使用机器翻译模型产生的伪数据或者使用 dummy 标签占位的方法来解决数据缺失的问题。伪数据生成法是利用已训练完成的神经机器翻译模型自动生成缺失语种的源句子，这种利用机器翻译生成的数据称为伪数据。利用伪数据生成法可以将所有语种的源句子准备完备，在数据完备的情况下，训练也就不存在任何问题了。占位法在训练的神经机器翻译模型时就设计了利用占位符来应对源句子缺失的情况，因此当所有源句子都存在的情况下，这个翻译模型可以产生更高质量的译文。

由于多源翻译系统是利用一个模型接受多个语种原文进行翻译的系统，因此，它可以在多种不同条件下部署。例如：当翻译延迟要求较高时，可以仅仅部署成一个双语模型，当翻译延迟要求一般时，可以部署成仅需要所有支持源语种中的几个语种的翻译模型，当对翻译延迟没有要求时，就可以部署成完整的多源翻译系统。

7.4.4 多源翻译的使用场景

多源翻译可以给译员提供一个相对较高质量的译文初稿，译员只

需要在多源翻译系统生成的译文之上做一些译后编辑工作就可以完成整个翻译工作。相对于让译员从零开始翻译原文，这种译后编辑的形式更省时省力，经济性也更好。多源翻译系统在原文有一个或多个不同语种的原文进行补充时可以产生质量更高的译文。多源机器翻译系统也可以用在多系统融合的翻译系统中，例如将统计机器翻译和神经机器翻译结合的场景中，多源机器翻译系统就能利用这两个系统的输出当成不同语种的输入，进一步提升翻译的质量。

总体来讲，多源翻译系统受到的关注较少，因为它往往需要多个步骤来完成翻译，所以翻译的实时性较低，无法像双语模型一样广泛用在各类翻译场景中。但当需要考虑如何实现一个灵活的多变的翻译系统时，多源翻译和多目标翻译应该与单源翻译-单目标翻译有着同样的地位。

7.5 领域适配问题

高质量平行语料一般只在某些特别的领域可以获得。无论是统计机器翻译还是神经网络机器翻译在低资源的、领域特定的翻译上效果都不好。如何充分利用领域外平行语料和领域内单语语料做领域内翻译是一个值得研究的问题，称为机器翻译的领域适配问题。

由于可以把每个领域看做一个语言，多语言神经网络机器翻译和神经网络机器翻译的领域适配之间存在很多相似性和共通方法。因此，类似于多语言神经网络机器翻译，使用领域外平行语料做领域适配时，多领域神经网络机器翻译和基于迁移学习的方法也用在了领域适配问题上。领域内单语语料的用法也以回译为主，类似于多语言神经网络机器翻译的伪平行语料生成。

MNMT 和领域适配之间也有很多不同。例如，基于枢轴的翻译在 MNMT 中很常见，但是在领域适配问题上不适用。不同领域之间一般总会有重合的单词，因此领域适配也没有零样本学习问题。以不

同风格写出领域内句子也并非难事，所以多源方法也不适用。另一方面，领域适配问题的一个常见解决方案是使用某种算法从领域外数据中选出与领域内数据相似的一部分，这种方法在 MNMT 中还没研究过。但是，随着跨语言句嵌入的发展，数据选择和实例权重调整等方法很可能在不久的未来用在 MNMT 上。

另外有一些工作尝试将 MNMT 和领域适配问题相关联，研究工作聚焦于在 MNMT 和领域适配问题上使用或改进微调方法。其中改进的微调方法可以是使用适配层，这只会引入少量的额外的参数。

本章编写人员：

张霄军、赖文、宗浩、杨曼芝

第 8 章 多模态智能信息处理

8.1 语音识别概述

如今随着人工智能和大数据的发展，机器也越来越有了学习能力，很多智能产品由之诞生，比如：自动驾驶、人脸识别、图像识别、自动语音识别技术等，在大家的不懈努力下，如今计算机已经非常聪明了，同时了充满了智慧，更具有学习和模仿人类的能力。当计算机出现后，计算机技术被广泛的应用的生活的每个领域，随之而来出现了大量的数据，其中语音识别技术备受大家的关注。

语音识别是一门交叉学科，其中涵盖了模式识别、信号处理、人工智能、信息论和概率论、语言学等知识。近年来，语音识别技术进入工业、军事、家庭、电子产品、汽车行业等领域种，并在其领域中发挥着不可或缺的作用。

同时随着 5G 技术的发展，语音识别也变得更加高效，目前的语音助手，智能语音音响，智能家居等设备的操作都是语音交互方式，人们也将会拜托键盘的束缚，让日常变得越来方便快捷，节省了大量的时间。同时低延迟的语音输入也会降低语音识别错误率，识别更加准确，语音识别的发展将会让全世界不同群体的人享受智能数字生活带给我们的便利，所以研究语音识别这项技术是很有实用价值的。

8.1.1 语音识别研究背景

语言是人类进行沟通交流的表达方式，是承载着人们的思维的符号体系，是人类社会主要的信息载体，而语音和文本是其两种主要的表现形式。某种意义上，语音信号是人与人之间传递信息，表达情绪的最直接的载体。语音信号中包含语义信息（语音直接表达的意思），说话人信息（说话人的年纪，性别）以及非语言学信息（各类背景音）等等。声波作为音频信号，和天线信号，视频信号等一样是非接触方式的传播，是不使用任何工具就能够接收和发出的天然的无线资源，这个特性在很多应用场景上给人们带来了极大的便利。语音识别也称

自动语音识别，主要任务是实现与机器进行语音交流，即让计算机‘理解’人类语言，由计算机完成音频信号到文本的转换。

语音识别技术从提出到实现经历了漫长的过程。1950 年图灵在发表《计算的机器与智能》，提出机器与人能否进行交流的问题。1952 年，贝尔研究所研发出第一个能识别出 10 个英文数字发音的音频识别系统。1970 年后，随着统计语言学的发展，德里克领导 IBM 华生实验室使用两个隐马尔可夫模型（HMM）实现更大规模的孤立词识别，奠定了之后很长一段时间内的语音识别框架：声学模型+语言模型。80 年代后，技术重点逐渐转移到非特定说话人的大词汇量识别，但是碍于语音信号资源的缺少以及计算机硬件的落后，语音信号处理技术停滞一段时间。1989 年，李开复主导开发 CMU SPHINX 工具，极大地推动了语音识别技术的工程化。

1973 年，我国科学院声学研究所开始计算机语音识别方向的研究，进入 80 年代之后，国内多个研究单位具备了研究语音技术的基本条件。在‘863’计划的支持下，语音识别被列为计算机系统中的一个专门的研究课题。2019 年，在北京互联网法院公开发布的《互联网技术司法应用白皮书》中，将语音识别技术列入十大典型技术应用。国内的很多科研单位纷纷投入研究语音识别技术，如中科大的语音及语言信息处理国家工程实验室，清华的语音技术中心等等。在过去几年里，语音识别技术逐渐应用到工业，家庭服务，通信，汽车，医疗，电子产品等各个方面。较为典型的是车载语音系统，智能客服，语音搜索，交互游戏等。

在过去近十年中，随着各类语音信号语料的极大丰富，计算机硬件的提高以及神经网络模型的发展，语音识别又迎来了新的高潮。端到端模型因其相较于概率统计模型不需要严格的发音词典和强制对齐处理，结构更为简单也更容易联合优化，语音识别系统逐渐从概率统计模型到端到端模型发展。维吾尔语作为一个资源匮乏语言，不仅

资源稀少，而且其资源质量低——地缘、文化、历史等导致出现了多种文字和个性化的拼写形式，即语音和文字都带有一定的噪音。本文围绕维吾尔语端到端语音识别展开研究。

8.1.2 语音识别研究现状

语音识别在生活中给人们带来了便利与智能，从上个世纪七十年代起就是机器学习，统计学习与信号处理等多门学科的一个重要的研究方向，国内外的研究机构在这个领域非常活跃，并且它在工业产业领域内也得到了广泛的发展与应用。

目前语音识别主要分为基于统计模型的传统框架识别：HMM-GMM/DNN-HMM 和基于神经网络的端到端识别框架：CTC/Attention。前者因其成熟的技术和扎实的理论目前在工业界占据着主导地位，后者随着计算机硬件以及神经网络的逐渐成熟有着巨大的发展潜力。

传统语音识别框架由 IBM 华生实验室在 1976 年发表，使用 HMM-GMM 框架实现了连续语音识别，奠定了语音识别之后几十年的基础。2011 年提出了 CD-DNN-HMM，首次使用深度神经网络代替高斯模型对发音状态和特征序列之间的发射概率进行建模。这之后，主流的神经网络如 CNN、RNN 等先后被用于进行声学模型建模。基于 HMM 框架的语音识别模型分为声学模型，发音词典与语言模型三个部分，声学模型对输入的音频特征和对应的音素序列进行建模，发音词典实现严格的单词与音素之间的映射关系，语音模型建立词与词之间上下文联系，最终将音频特征序列转换为文本序列。

在基于 HMM 的语音识别框架中，声学模型的质量决定了识别系统的性能，音频信号作为 HMM 中的可观测序列，声学特征作为隐藏序列，通过 HMM，语音中的每一帧音频都对应着一个使用 HMM 进行建模的发音状态，所有可能的发音状态的组合被表示为一个巨大的有向图。在解码时使用维特比算法在这个有向图中搜索出最优路径，

路径所代表的内容就是语音识别的结果。虽然现在有工具箱如SPHINX，KALDI 非常完备地集成了基于 HMM 的识别框架，但是依然面临着几个难题：

（1）构建完善严格的发音词典。

（2）训练模块多，流程线较长，难以全局优化。

基于神经网络的端到端语音识别框架在 2006 年之后取得进展，Graves A 在中提出链接时序分类，2016 年，谷歌在中提出基于注意力机制的编码解码识别网络 LAS（litsen，attend and spell），2018 年，谷歌使用多头注意力机制代替单一的注意力机制，在数十万小时语料的支撑下取得 SOTA 结果。2019 年谷歌提出 conformer 模型，增加了 transformer 模型对于局部信息的捕获能力。2018 年，脸书提出 wav2letter，2020 年提出 wav2vec2.0，使用无标注的音频数据训练预训练模型，在少量的有标签数据上进行迁移学习，取得了 SOTA 结果。2014 年，百度提出 DeepSpeech 框架，阿里在论文提出 DFSMN。

与基于 HMM 的框架相比，端到端语音识别框架 1）将识别流程整合到一起，极大简化了识别流程，并且方便选择与评估标准强相关的函数训练模型，找到全局最优解。2）不依赖于严格发音词典的构建和强制的数据对齐，降低了语音识别的门槛。端到端语音识别框架虽然取得了很不错的结果，但是：

（1）对训练音频数量的依赖，越多的音频取得的预测结果越好，相反，少量的音频数据不能取得使用的性能。

（2）对计算力的依赖，神经网络模型和大量的训练数据需要强大的计算力做载体。

2009 年，Dan Povey 主导整理和研发语音技术工具 kaldi，kaldi 使用 C++和 shell 脚本，集成复杂的语音识别模块并将其链路化，成为近十年内甚至更长时间内的主流语音识别技术工具。2018 年，shigeki 等人创建端到端语音识别工具箱 espnet，集成 Google 提出的

基于注意力机制的端到端语音识别模型（LAS）与 Alex Graves 提出的基于链接时序分类的语音识别模型（RNN-CTC）。2019 年，facebook 提出 wav2word，次年提出 wav2vec，并在 fairseq 中发布 wav2vec 的工具。

8.1.3 低资源语言识别

随着计算机使用范围的越来越普遍和计算机技术和信息处理的发展，全球经济的形成和市场一体化，以及近年来中国的快速发展，对语音识别的需求也越来越旺盛。此外，国内外很多语言的语音及语言处理方面的研究取得了一定的成果，尤其是国家在政策、资助等方面给予的大力支持下，我国各民族文化、语言、文字等在民族信息化建设中得到了极大地改善与发展。但对属于黏着性的低资源少数民族语言来说，实现最优的多种低资源语言文本及语音单元并进行应用处在一片空白，而且对多语言信息处理及语音识别具有非常重要的研究意义，是新的研究内容之一。

单个少数民族语言的信息处理工作已经有了一定的进展，但没能在通用性和低资源语言方面得到推广应。因此，近几十年以来语音识别研究在世界范围内蓬勃发展起来，尤其是低资源语音识别研究的需求越来越称为热门了。低资源多语言信息处理指的是用计算机来处理维吾尔语、哈萨克语、柯尔克语等属于黏着性语言的文本，语音，形态，语义等信息最终为少数民族语言处理提供统一的、支持多种语言的、用户界面的集成信息处理软件环境。

在国家的大力支持下，维-哈-柯等少数民族语言的信息处理研究得到很好的发展。但是，由于资源匮乏，标准化工作跟不上等原因，发展比较缓慢。传统语音识别系统框架由声学模型（AM）和语言模型（LM）两个核心模块构成。在国家的大力支持下，少数民族语言，尤其是维吾尔语言语音识别研究有了一定的研究基础。神经网络模型在更小的语料资源上的高效的学习能力给资源匮乏语言的信息处理

研究带来了新的机遇。

通常为了提高集成软件环境的通用性，尽量减少语言相关的人工工作量，主要思路是将规则的收集、学习、训练、及切分等工作凝练成一个统一的软件框架，学习和训练部分与语言无关。每个语言只提供若干句子的词-词素平行语料库，和词缀库（闭集合）。通用软件直接从平行语料库中自动学习规则、语音和谐、上下文等信息，并对测试句子提供所有可能的切分序列结果。然后由独立的统计模型对该词素序列进行排序。

8.1.4 语音识别难点

目前，语音识别还是存在许多技术需要改进的地方：

（1）模型的改进需要做进一步工作。语音识别的声学模型和语言模型这两个地方运用到模型的地方，虽然增大语料可以提高语音识别准确率，但是数据的收集需要很大的工作量，因此采用和改进模型也是需要突破的地方。

（2）在研究语音识别时，在研究中一般是纯净的语音的数据，然而，在现实生活中都会参杂一些噪音，所以提高语音的强健性是很有必要的，能在噪声环境下准确的识别出语音也是一项艰巨的任务。

（3）语音自适应技术有待提高，说话人在用不同性别、方言、口音说话时，由于语言模型训练有限，导致识别率低，为了能满足更多说话人声线特征，研究说话人自适应技术也是推动语音识别发展的方向。

（4）多语言的混合模型也是今后研究的热点。目前在单语言说话人的识别效果挺好的，但是换另外一种语言说话，就会识别不出，所以将多个语言模型的融入，就不受单语种的影响，可以多种语言自由切换自动识别。

8.2 语音识别技术

8.2.1 传统 ASR 系统框架

自动语音识别(ASR)是将一段输入语音中承载的信息转换成文字的过程。从原始语音数据中提取声学特征并通过统计训练得到声学模型，作为声学单元的模板，与语言模型结合，经过解码处理得到相应的识别结果。传统语音识别系统框架如图 8-1 所示，系统由解码器、声学模型、语言模型、发音字典等多个独立部分组成并独立优化。

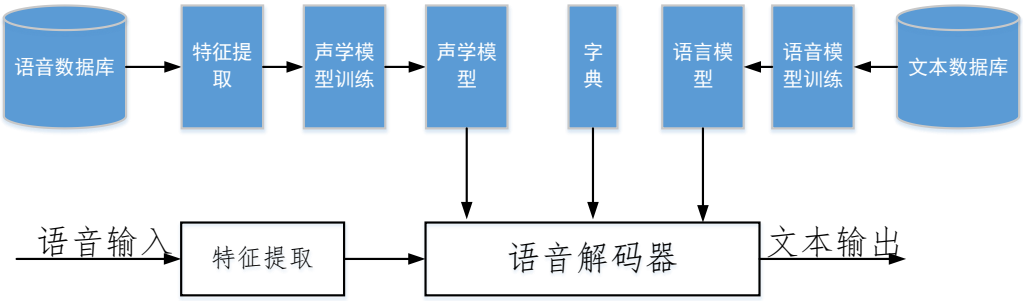


图 8-1 语音识别系统框架示意图

8.2.2 语音信号的特征表示

系统输入特征质量的好坏一定程度上直接影响到识别结果，所以它被认为在语音识别系统中是一个重要的环节。特征提取过程是将压缩输入语音信号的幅度，在这个过程中语音信号的功率造成不会受到损害。语音信号特征提取技术有很多种类。常用的语音特征有线性预测系数、梅尔频率倒谱系数(MFCC)和基于滤波器组的 Fbank 特征等。

其中，LPC 是根据人的发声机制的特点提出来的方法。MFCC 是根据人的听觉机制的特点提出来的方法。主要使用人耳听觉标度的 Mel 标度。它是提取频谱特征的最普遍和最常用的方法之一。在基于 GMM-HMM 的语音识别系统中也采用 MFCC 特征提取技术。本文也用 MFCC 特征作为输入特征序列。该特征提取技术的过程如图 8-2 所示。

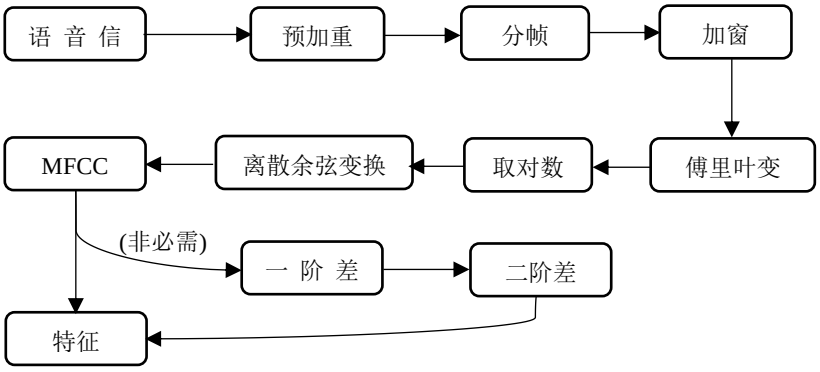


图 8-2 MFCC 特征提取过程示意图

1. 预加重

对语音信号而言，信号中承载的能量主要分布在信号的低频部分，随着信号频率的增高，信号的输出信噪比会明显下降，从而导致高频信号在传输过程中衰弱，最后给语音信号的整普段质量带来严重的影响。为了解决上述问题，对语音信号作预加重，一方面平衡信号的高低频率，另一方面提高信号中高频部分的分辨率，最后获得频谱更加平淡的信号。预加重操作一般用有限长单位冲激响应(Finite Impulse Response, FIR)滤波器来完成：

$$y(n) = x(n) - ax(n-1)$$

式中a 为预加重系数，取值范围一般在 0.9~1 内，通常取值为 a =0.98；$x(n)$是在第 n 时刻的采样值，$y(n)$为预加重结果。

2. 分帧和加窗

语音信号为非平稳信号，它的统计属性是随着时间变化的，这种不稳定性主要有发声器官和周围环境所引起的。以汉语为例，一句话中包含很多声母和韵母，不同的拼音，发音的特点很明显是不一样的。但是，研究表明，在 10 毫秒至 30 毫秒(ms)范围之内的语音信号又具有短时平稳的属性。比如汉语里一个声母或者韵母，往往只会持续几十到几百毫秒，在这一个发音单元里，语音信号表现出明显的稳定性和规律性。因此，把平稳过程的理论引入到处理语音信号处理的领域。

分帧将把每一段语音信号划分为许多个短时平稳的语音段，每一条语音段被称一个帧，帧长一般 25ms 左右。然后是帧移，这将两个帧之间形成一定程度的重叠，使得帧与帧之间的过渡变得平滑自然，长度一般 10ms 左右。图 8-3 是帧长与帧移示意图。

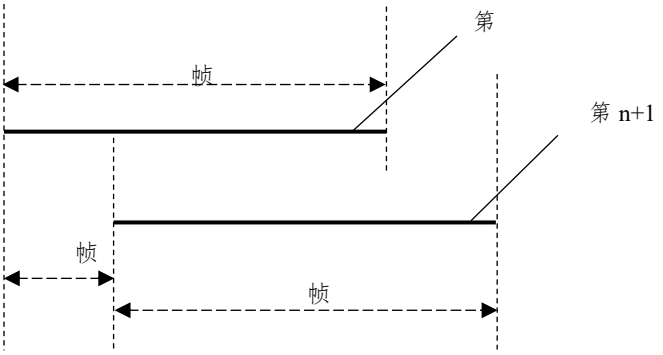


图 8-3 帧长与帧移示意图

分帧的过程实际上是在时域上用一个窗函数和原始信号进行相乘—信号加窗。加窗是为了减少语音信号的帧起始和结束位置信号的不连续性问题以及便于进行傅立叶展开，使得每一帧信号在全局上更加连续。一般采用汉明窗进行加窗，是语音信号预处理时最常用的窗函数。它的主瓣比较宽，旁瓣衰减大，具备了平滑的低通特性，能在较高的程度上反应短时平稳信号的频率特性，其定义如下公式所示：

$$W(n)=\begin{cases}(1-a)-a*\cos\left[\dfrac{2\pi n}{N-1}\right], & (0<n<N-1)\\ 0, & \text{其他}\end{cases}$$

式中a 是可调常数，通常取值为 0.46。

3. 语音特征提取

加窗分帧之后的语音信号依然是属于时域信号，其体现的信息较少。为了观察到信号的能量分布情况，对每一帧时域信号进行快速傅

里叶变换获得它们的频谱，然后取模获得语音信号的功率谱密度：

$$X_c(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-\frac{j2\pi k}{N}\right) \quad (0 \le k \le N)$$

式中$x(n)$ 加窗分帧后的语音信号，N为快速傅里叶变换的点数。随后采用一组梅尔尺度的三角形滤波器组，对已获得的语音信号功率谱进行平滑操作。平滑操作不仅消除谐波效应，又能降低运算量和复杂度。滤波器组的频率响应定义为如下公式形式：

$$H_m(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m) \le k \le f(m+1) \\ 0, & f(m-1) \ge k \text{ 或 } k \ge f(m+1) \end{cases}$$

其中 $\sum_{m=0}^{M-1} H_m(k) = 1$ 。设计一个由 M 个三角形滤波器组成的滤波器组，M 通常取值为 24，每一个滤波器的中心频率为 f(m)，该三角形滤波器组如图 8-4 所示：

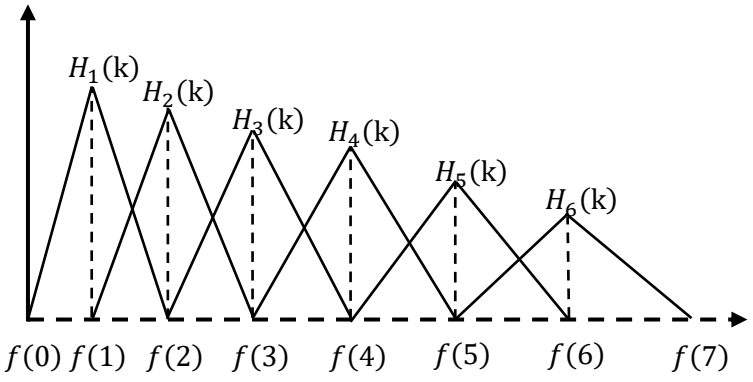


图 8-4 三角形滤波器组

从线性频率到 Mel 非线性频率的转换由如下公式来进行：

$$Mel(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right)$$

式中 f 表示线性频率，单位为 Hz 。然后计算每个滤波器输出的对数能量，数学表达式如下所示：

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X_c(k)|^2 H_m(k)\right) \quad (0 < m < M)$$

完成了计算对数能量的步骤，我们得到梅尔频率滤波特征(Fbank)序列。对 Fbank 特征再进行离散余弦变换(DCT)操作可得 MFCC 特征序列：

$$C(n) = \ln\left(\sum_{k=0}^{N-1} S(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right)\right) \quad (n = 1,2 \ldots L)$$

上式中 M 是指三角形滤波器的个数。

4. 一、二阶差分

MFCC 特征只反映了语音参数的静态特征，如果涉及到动态特征序列还可以计算出静态特征的一阶、二阶差分以及信号能量的一、二阶差分作为系统输入特征序列，一阶差分（Delta, Δ），类比速度，最简单的一阶差分计算方法：

$$\Delta(n) = \frac{C(n+1) - C(n-1)}{2}$$

二阶差分（Delta Delta, $\Delta\Delta$）类比加速度，简单计算方法：

$$\Delta\Delta(n) = \frac{\Delta(n+1) - \Delta(n-1)}{2}$$

计算能量：

$$E = \sum_{n=0}^{N-1} |x(n)|^2$$

MFCC 是语音识别任务中最常用的语音特征之一，此特征对基于 GMM-HMM 的语音识别系统比较合适。对基于 DNN、端到端等语音

识别系统来说，网络的输入需要相关性较强的特征，而 MFCC 特征相关性较弱，不是很适合应用基于神经网络模型的系统中。因此，经常采用特征相关性较强的 Fbank 特征。MFCC 和 Fbank 特征计算步骤只有一步之差，对 Fbank 特征执行一次离散余弦变换便可得 MFCC 特征。下面对 MFCC 和 Fbank 进行简单的对比：

1）MFCC 特征在 Fbank 特征的基础上获得。因此相比 Fbank 特征，MFCC 需要更大的计算量。

2）特征区分度方面，Fbank 特征具备了相关性较高的特点，因为相邻滤波器组有重叠的部分；而 MFCC 特征具备了更好判别读的特点，这也是在大多数语音识别系统中使用 MFCC 的原因。

3）基于神经网络的语音识别模型能够更好地利用相关性高的特征，能够更有效地降低词错误率 WER。

4）与 MFCC 相比，Fbank 特征包含更多的语音特征信息。因此在 Fbank 基础上用 DNN、端到端模型建立声学模型时系统学习速度更快，稳定性更好。本文第四、五章建立的语音识别系统中用的都是相关更好的 Fbank 特征。

8.2.3 声学模型

声学模型描述的是词的发音信息，也是语音识别核心模块之一。自动语音识别(ASR)主要研究从长度为 L 的语音特征序列，假设 $X = \{x_L \in V | t = 1, \dots, L\}$，到长度为 N 的单词序列，即 $W = \{w_n \in W | n = 1, \dots, N\}$，的序列映射问题。换句话说，假设 X 是指给定语音特征序列(观测序列)，W 表示单词序列，则最有可能的单词序列 W^* 可以有以下公式计算出来：

$$\widehat{\mathrm{W}} = \arg \max_{\mathrm{W} \in \mathrm{W}*} \mathrm{p}(\mathrm{W}|\mathrm{X})$$

于是，如何计算出后验概率 $p(W|X)$ 成为语音识别最主要的问题。通常采用贝叶斯定理来解决后验概率的问题，$p(W|X)$可以分解为以下形式:

$$p(W|X) = \frac{p(X|W)p(W)}{p(X)}$$
$$\propto p(X|W)P(W)$$

然后，

$$\arg\max_{W \in W*} p(W|X)$$
$$= \arg\max_{W \in W*} \sum p(X|W)\, p(W)$$

公式中，$p(W|X)$ 和 $p(W)$，分别是声学模型和语言模型。声学模型，$p(X|W)$,通过使用概率链规则和条件独立性质进一步分解，如下所示：

$$p(X|W) = \prod_{t=1}^{T} p(x_t|x_t, \ldots, x_{t-1}, W)$$

$$\approx \prod_{t=1}^{T} p(x_t|w_t) \propto \prod_{t=1}^{T} \frac{p(w_t|x_t)}{p(w_t)}$$

传统声学模型基本上为 GMM-HMM，在之后技术的持续延伸和拓展下，出现了 DNN、DNN-HMM 和 CNN 等多元化的声学模型。声学模型多元化发展，在一定程度上为语音识别技术突破提供了有力保障。

8.2.3.1 GMM-HMM 声学模型

分析早期语音识别声学模型能够发现，GMM-HMM 声学模型在实践中得到了普遍运用这一类声学模型体现出结构单一的显著特征，且能够完成高效的区分度训练。结合高斯混合模型进一步拓展构建出的隐马尔可夫模型，能够对具有直观观察特性的高斯过程以及体现出显著隐蔽性的马尔可夫过程加以有效随机组合。其内部参数集合包括了单一的状态先验概率向量、状态转移矩阵、以及 GMM 参数。HMM 在进行功能发挥期间对“上下文依赖状态”进行了合理运用，促使各状态语音特征向量能够表现出一致性的统计特性。因此能够实现对细

节特性的高效反馈。然而模型基于条件独立以及分段平稳等条件的基础之上。所以在之后的研究中，相关语音识别学者可以围绕能够进行多类真实语音时域动态属性反映的统计模型，展开相关研究构建工作。

GMM-HMM 声学模型的构建，为语音识别技术飞速发展提供了有效的推动力。在模型运用期间，首先需要展开诸多高斯分量的加权处理，促使整个声学特征空间分布能够在合理科学的手段下完成模拟调控。之后再基于无监督的 EM 算法对系列数据集展开声学模型的相关训练。与此同时，在国家的大力支持下，低资源语言语音识别研究也有了一定的基础。

8.2.3.2 DNN-HMM 声学模型

DNN 难以直接结合语音信号完成模型构建，这主要是由于前者为时序连续信号，但该模型中对于信号方面，往往需要进行固定大小的输入。1990-1991 年部分专家学者通过进一步研究构造出了多层感知机-隐马尔科夫混合模型（MLP-HMM）。值得注意的是前期展开的混合模型语音识别研究基本上是不体现上下文关联性的音素状态，且只能够局限于小词汇量层面的语音识别。2011 年，George E.Dahl, Dong Yu 等研究人员设计了深度神经网络-隐马尔科夫混合模型 DNN-HMM。HMM 主要可以结合语音信号序列属性完，成对应的模型构建，进而实现对整个信号动态变化的模拟分析；DNN 为是基于 HMM 发射概率（emission probability）要素提取下进行的建模，能够就聚类后三音素状态的似然度完成建模操作，并进一步剖析出观察特征的概率参数。当声学观察特征确定之后，DNN 输出层节点可以实现对连续密度 HMM 状态后验概率准确有效估计。其作用原理为：首先系统会获取到语音信号的声学特征，随后在 DNN 支持下进一步将特征同隐层空间完成对应映射，最终将最后一层隐层借助 softmax 函数实现声学特征向状态空间的映射。这种持续的非线性变换结构，促使 DNN 在复杂数据描述方面表现出比 GMM 更为优异的性能。另

一方面，也可以将 Viterbi 算法运用到 DNN-HMM 训练期间，解码过程也表现出高效性特征。

8.2.3.3 CNN-HMM 声学模型

卷积神经网络(Convolution Neural Network，CNN)为前馈神经网络，是通过有监督学习应用，加以构建的数学模型，该类神经网络包括输入层、多个卷积层（convolutional layers）、池化层(pooling layers)等多个层面。各层彼此之间交替出现促使整个网络的顺利构建。前端可以发挥对特征提取，后端具有诸多全连接层(full-connected layers)，主要发挥对获取的局部特征加以整体的整合和变换功能，随着任务动态的持续变化，网络最终输出情况也会随之产生一定的调整改变。卷积神经网络通过调整自身参数便能够很好地拟合任意数据分布，体现出显著的特征学习能力，能够从大规模原始语音数据内进行有用数据的针对性提取，并借助多个隐层完成持续的空间变换，最终在 softmax 层输出支持下，映射到状态空间。

CNN 结构如图 8-5 所示，受特殊网络结构的支撑，CNN 在输入特征方面体现出显著的稳定性。

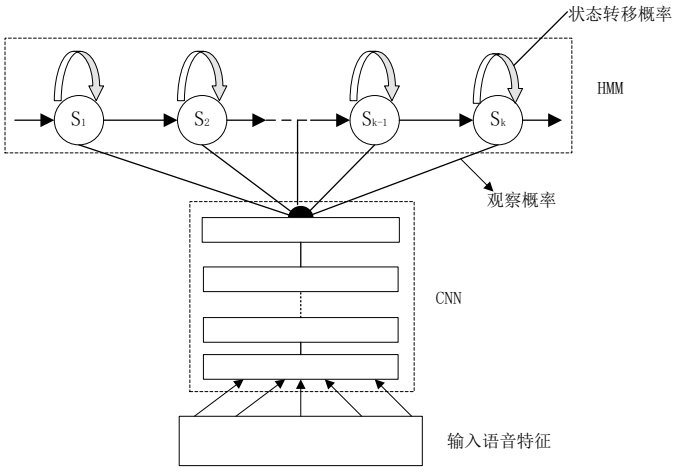


图 8-5 CNN-HMM 模型结构图

CNN-HMM 和 GMM-HMM 声学模型训练时，GMM 和 CNN 模

型的任务是获取所有语音帧与 HMM 状态之间的对应关系。可以先对 GMM-HMM 声学模型展开相关的训练，选取以此掌握到系列状态类别和状态量的转移概率参数。随后结合初始状态确定 CNN 输出层神经元数目并输出标签， 开始第二次 CNN 的有监督训练。CNN 的输出层可以用 Softmax 层来输出各个状态，并训练 HMM 状态的后验概率。在解码期间，通过 CNN 获取到的状态后验概率，要基于贝叶斯公式进一步转化为可以有效兼容 HMM 模型的观察概率。

8.2.3.4 端到端声学模型

端到端的语音识别系统完全不同于使用 HMM 结构进行语音识别的方法，它在训练过程中自动学习并优化语音信息和标注序列的对应关系，而不需要在网络进行训练前先得到帧对齐，减少了对齐对声学模型性能的影响，使语音识别系统更加直观。

1. 基于 CTC 的端到端语音识别

链接时序分类（connectionist temporal classification）是一种用于解决时序分类问题的损失函数，由 Graves 等人在提出，在 2013 年被用用于语音识别任务。不同于基于 HMM 的语音识别框架使用交叉熵作为损失函数，CTC 通过使用最大似然标准直接优化输入序列和输出序列的似然来解决时序分类问题，基于 CTC 的神经网络训练准则称为 CTC 准则。

CTC 损失函数如公式：

$$\mathrm{Loss(S)} = -\sum_{(\mathrm{x},\mathrm{z})\in \mathrm{S}} \ln(\mathrm{p(z|x)})$$

其中 z 为目标序列，S 为训练集空间。

CTC 不需要使用 HMM 模型，同时也解决了 DNN-HMM 需要对数据进行强制对齐的问题。在众多不同的序列分类任务（手写体识别，语音识别）上的出众结果表明，具有较强上下文能力的神经网络模型，（如双向长短时记忆神经网络，transformer 神经网络）与 CTC 损失

函数进行组合，能够得到比基于 HMM 的识别模型更为优异的结果。

基于 CTC 的端到端语音识别框架如图 8-6：

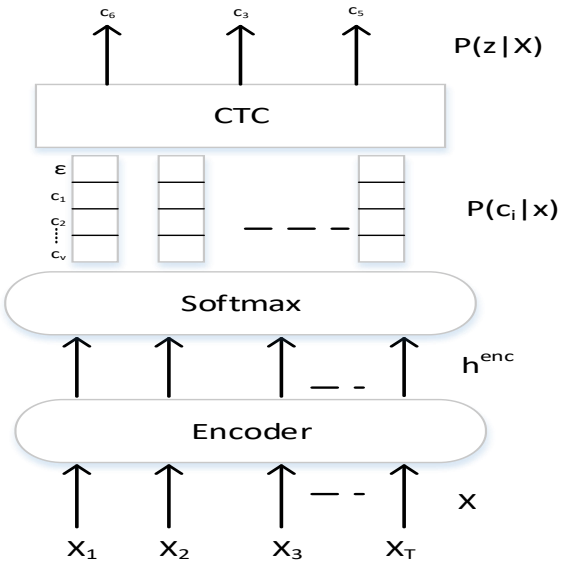


图 8-6 基于 CTC 的识别框架图

特征序列 X=（X1，X2…XT）经过神经网络模型进行编码得到 henc=（h1，h2…hT）,输出端使用 softmax 层进行线性处理之后得到后验矩阵 y=（y1,y2…yT）,之后使用 CTC 损失反向传播训练网络。

CTC 引入 blank 字符解决变长映射的问题:语音序列的学习任务是多对多的映射关系，n 帧音频输入对应的是长度为 m 的字符序列或者音素序列（m<=n）,CTC 提出引入 blank 字符解决这类变长映射问题。blank 字符的具体作用是标记静音和分隔标签。经过神经网络进行编码之后得到是大小为 T*N（N 为标签表的大小，在语音识别中也就是识别单元的多少）的后验矩阵，这个矩阵表示的是从 1 时刻到 T 时刻每个识别单元的后验概率。会出现连续多帧音频映射到同一个识别单元或映射到静音标签的情况。

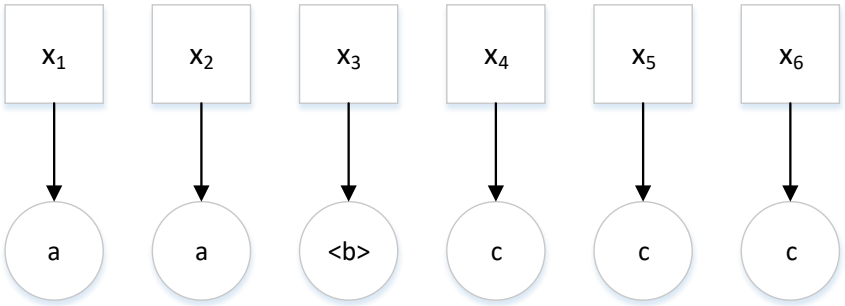


图 8-7 多个连续时刻的特征映射到相同标签或静音

CTC 中的 B 变换算法与 π 路径:通过神经网络将音频特征映射到识别单元的后验矩阵之后，在后验矩阵代表的识别单元序列中搜索出最优的词序列。B 算法与 π 路径解决了不同的音素序列映射到相同词的问题，如：cc<b>aa<b>t，cc<b><b>a<b>t<b>，cccc<b>aaaaaaa<b>ttttttt 三个序列都映射到 cat。我们用 π 表示一条由识别单元表中元素组成的长度为 T 的路径，路径经过 B 变换之后得到 z。

B 变换算法：

1)扩展原始标签表：z’ = z U blank;

2)对输出的字符串进行 B 变换 z=B（z’），将重复的字符合并为一个，去除 blank 字符得到最终的单词。

CTC 损失函数:CTC 使用最大似然标准作为损失函数。

存在不同的路径 π 1, π 2, π 3... π m 过 B 变换之后能够得到 z。所以损失函数可以得到如下：

$$p(z|x) = \sum_{\pi \in B^{-1}(z)} p(\pi|x)$$

B-1 是 z 到由 B 变换可得到 z 的全部路径集合的映射函数。CTC 输出的概率是相对于输入条件独立的，所以输入为 x 时路径 π =（ π 1， π 2， π 3... π T） 的概率为：

$$p(\pi|x) = \prod_{t=1}^{T} y^{t}_{\pi_t}, \forall \pi \in z'$$

p(z|x)为所有的路径概率之和，损失函数即：

$$\mathrm{Loss(S)} = -\sum_{(x,z)\in S} \ln\Big(\sum_{\pi \in B^{-1}(z)} \prod_{t=1}^{T} y^{t}_{\pi_t}\Big), \forall \pi \in z'$$

CTC 采用动态规划的方法使用前向后向算法有效地训练网络，使用前向后向算法计算前向概率 α 和后向概率 β：

$$\alpha_t(s) = \begin{cases} \Big(\alpha_{t-1}(s) + \alpha_{t-1}(s-1)\Big) y^{t}_{z'_s}, z'_s = \text{blank or } z'_{s-2} = z'_s \\ \Big(\alpha_{t-1}(s) + \alpha_{t-1}(s-1) + \alpha_{t-2}(s-2)\Big) y^{t}_{z'_s}, \text{otherwise} \end{cases}$$

$$\beta_t(s) = \begin{cases} \Big(\beta_{t+1}(s) + \beta_{t+1}(s+1)\Big) y^{t}_{z'_s}, z'_s = \text{blank or } z'_{s+2} = z'_s \\ \Big(\beta_{t+1}(s) + \beta_{t+1}(s+1) + \beta_{t+2}(s+2)\Big) y^{t}_{z'_s}, \text{otherwise} \end{cases}$$

s 表示该时刻经过第 s 个节点。

在任意时刻 t，利用前向概率和后向概率计算 CTC 损失：

$$\mathrm{p(z|x)} = \sum_{s=1}^{|z'|} \frac{\partial_t(s)\,\beta_t(s)}{y^{t}_{z'_s}}$$

$$-\ln(\mathrm{p(z|x)}) = -\ln\Big(\sum_{s=1}^{|z'|} \frac{\partial_t(s)\,\beta_t(s)}{y^{t}_{z'_s}}\Big)$$

求出$\frac{\partial \mathrm{p(z|x)}}{\partial y^{k}_{t}}$之后通过反向传播方法求$\frac{\partial \mathrm{p(z|x)}}{\partial \omega}$训练神经网络参数 ω。

CTC 的局限：CTC 在进行概率计算的过程中假设每个识别标签是相互独立的，而实际上每一时刻的标签都与上下文有着依赖关系，这个问题可以使用语言模型（language model，LM）来解决。

2. 基于 attention 的端到端语音识别

2014 年，google 的 mind 团队在 RNN 模型上使用了 attention 机

制来进行图像分类，之后 Bahdanau 等人将 attention 机制用于机器翻译任务上。2015 年，google 的 Navdeep Jaitly 等人提出编码-解码（encoder-decoder）模型用于语音识别，2018 年 google 使用多头注意力机制代替 LAS 模型中的注意力机制，在语音识别任务上取得了 SOTA 结果。

基于 Attention 的语音识别系统，系统分为三个模块：编码器，注意力模块，解码器。编码-解码框架在此之前已经在自然语言处理领域中的机器翻译方向被用来解决输入和输出都是不定长序列的序列预测问题。编码器将不定长的输入编码为一个定长的序列，并将序列信息映射为高维的特征，通常使用能够捕获上下文信息的神经网络。解码器将编码后的固定向量再转换为输出序列。编码-解码框架如图 8-8：

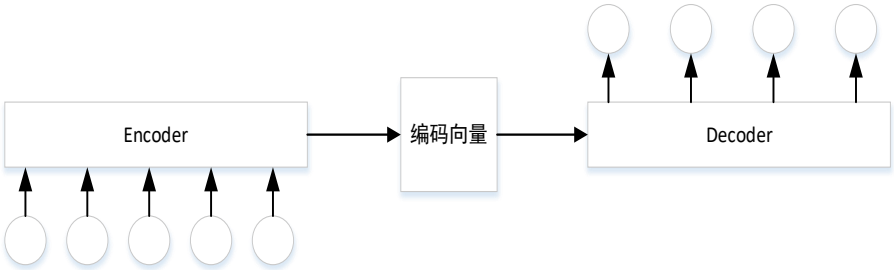


图 8-8 编码-解码框架

注意力机制模块本质来源于人类的视觉注意力机制-根据接收到的需求在观察中注意到场景中自己所想的某一部分。注意力模型通过分配权重让输入序列中的每个元素聚焦到其他元素上，以此来捕获上下文信息。基于 attention 的端到端语音识别同样不需要强制对齐操作，它使用 attention 模块实现软对齐（soft alignment），这也使编码器的编码输出不在必须是一个固定长度的矢量。

基于 attention 的端到端语音识别框架如图 8-9：

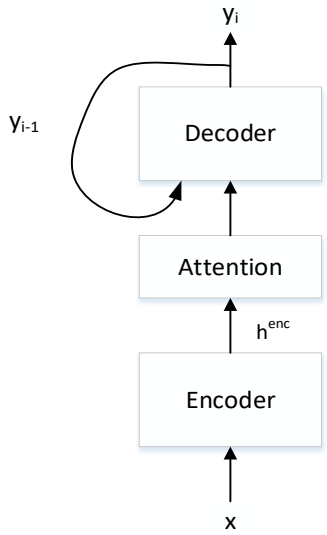


图 8-9 LAS 端到端模型

编码器将输入特征序列 X 编码为更高维的特征表示 h=(h1,h2…hU),U<=T：

$$h = \text{Encoder}(X)$$

解码器将编码输出 h 以及 i 时刻之前的序列解码为 yi：

$$p(y_i|x, y_{<i}) = \text{Decoder}(y_{<i}, h)$$

其中 y=（<sos>，y1，y2，…，ys，<eos>）,yi∈{识别单元表，<common>,<period>,<apostrophc>,<unk>},<sos>和<eos>分别是开始标识符和结束标识符。

将解码器（以 RNN 为例）展开如图 8-10：

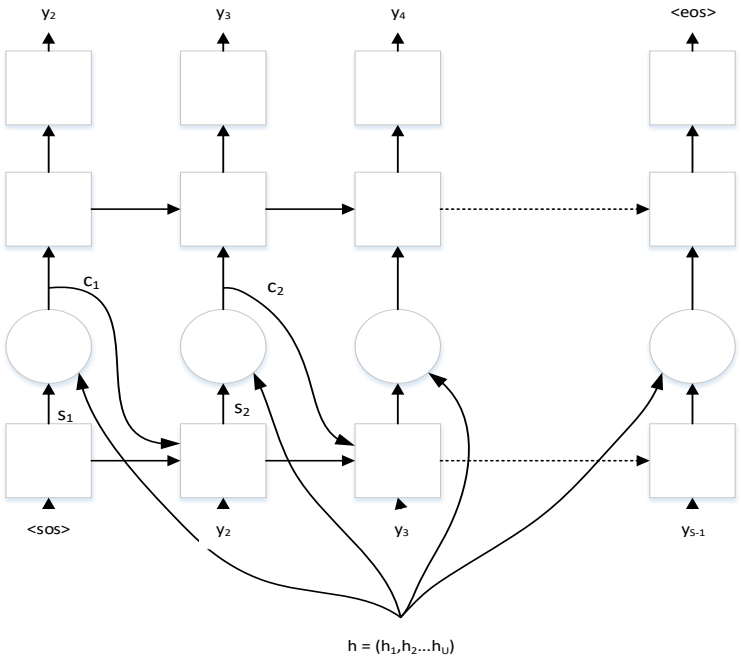


图 8-10 解码器展开结构图

$$s_i = \mathrm{RNN}(s_{i-1}, y_{i-1}, c_{i-1})$$

$$c_i = \mathrm{AttentionContext}(s_i, h)$$

模型在训练过程中最大化真实标签的概率，但是在预测过程中，由于真实标签的缺失可能会影响最终的识别结果，所以在训练过程中 LAS 会从之前的字符分布中采样，并且将它作为发射状态用于预测下一状态，如下：

$$\max_{\theta} \sum_{i} \log P(y_i | x, y^*_{<i}; \theta)$$

其中$y^*_{<i}$是真实标签或者 CharacterDistribution（si，ci）

LAS 的缺陷：LAS 使用 attention 模型捕获长距离上下文信息，且相较于 CTC 模型没有使用条件独立假设，所以在不使用语言模型的情况下，能够得到更高的 CER。但是在实际的语音识别任务中，注意力机制中的对齐很容易被噪声破坏，而且对于较长的音频，需要手动设置窗口限制注意力模型的探索区域。而且复杂的网络结构冷启

动对于小语料来说并不友好。

8.2.4 语言模型

结合实践情况来看，在进行语音识别期间，往往在解码语音信号时会产生对应的一组候选词序列。语言模型（Language Model，LM）基于对候选词序列出现概率和可能性的分析，进一步对各个选词序列完成概率得分判断，之后再结合最终的识别结果，将综合分数最高的词序列加以呈现。

8.2.4.1 N-gram 语言模型

N-gram 语言模型在语音识别领域广泛应用，最流行的统计语言模型，构建模型简单，但体积大。N-gram 语言模型以独立性假设为条件建立的，即第 N 个单词的出现仅仅与它前面的 N-1 个单词有关。其中，N=1 的情况称为一元方法（uni-gram），N=2 的情况称为二元方法（bi-gram），N=3 的情况称为三元方法（tri-gram）。tri-gram 语言模型是目前运用范围较为广泛的语言模型。

语言模型表现出来的性能差异主要可以结合交叉熵（cross-entropy）和困惑度（perplexity，PPL）加以体现。测试集 *T* 主要由一系列句子 *（t1,t2,...,ti）*，概率 *P(T)*对应于 *T* 内所有据此出现的概率乘积。若结合*-log2P(T)*个比特位展开对文本 *T* 相关编码操作，则对应的交叉熵 $H_P(T)$可表示为：

$$H_P(T)=-\frac{1}{W_T}\log_2 P(T)$$

其中 W_T 是文本 T 的长度，词汇总数。

困惑度 PPL 简单来说就是测试集 T 内每个单一词汇概率几何平均数的倒数，困惑度与交叉熵存在以下关联性：

$$PPL(T)=2^{H_P(T)}$$

随着交叉熵和困惑度降低，语言模型表现出的识别效果越显著，系统识别文本的性能也越高。

8.2.4.2 RNN 语言模型

2012 年 Mikolov 将 RNN 引入语言建模后，RNN 语言模型在相关文字输入以及翻译方面得到了大规模应用。且形成的应用效果显著优于 N-gram 语言模型。从 Mikolov 做的实验所得数据可知：RNN 语言模型能够进行词错误率（WER）的高效控制，比 N-gram 语言模型具有更高的准确率。而且，RNN 的隐层可以记忆能记忆更多的上下文信息。因此，本文针对语言模型构建，最后确定选择实践应用价值更高的 RNN 语言模型。该模型的网络结构如图 8-11 所示。

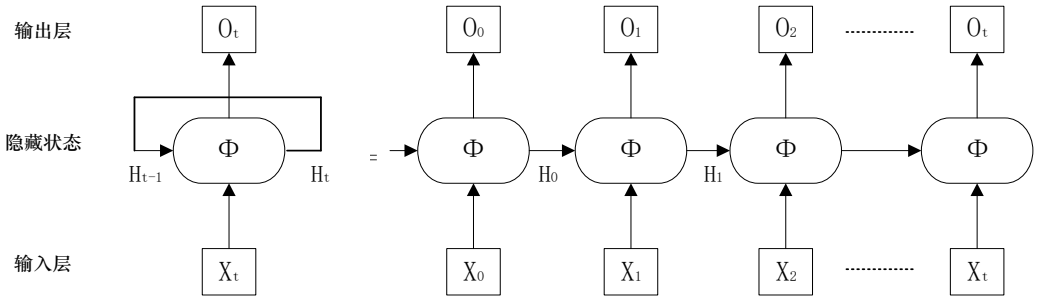


图 8-11 RNN 的网络结构图

以上图中左边的是 RNN 的基本结构图，观察图可知，RNN 主要由输入层、隐层和输出层等三个网络层构成的[40]。右边的是按时间序列展开之后的展开图，简单的说，如果我们需要分析包含 5 个词的句子，将 RNN 模型可以展开为 5 层神经网络，即展开每次为一层的形式。

② Xt 是 t 时刻的输入（比如，X0 是 t=0 时刻的输入）；

② Ht 是 t 时刻的隐藏状态，即循环神经网络的记忆，它由 Xt 和 Ht 共同决定。计算公式为如下：

$$H_t = f(U \cdot X_t + W \cdot H_{t-1} + \beta)$$

其中 U、W 和 β 均为网络参数，f 为激活函数（tanh 函数或者 ReLU

函数），以上公式的计算是循环的。

③ Ot 是 t 时刻的输出，输出大小取决于磨课目前所处时刻的隐藏状态 Ht（记忆），即，将所有时刻的输出概率进行相加，则最终结果为 1；

$$O_t = soft\,max(V \cdot H_t + \eta)$$

其中，V 和 η 也均为网络参数，激活函数为 softmax。

④ Φ 是 RNN 模型；

上面描述的是经典 RNN 的运算过程。

8.2.4.3 神经网络语言模型

Y Bengio 等人于 2003 年最早实现神经网络语言模型（Neural Network Language Model，NNLM），对于文本序列 w，NNLM 实现的目标是根据前 n 个文本符号出现的概率预测当前文本符号的概率：

$$f(w_l|w_{l-1}\dots w_{l-N+1}) = p(w_l|w_{l-1}\dots w_{l-N+1}),\ f>0 \text{ 且 } \sum_{i=1}^{V} f = 1$$

其中 v 是词表大小。通过式（）中的条件概率的乘积，得到整个文本序列的联合概率。实际部署中又分为：文本特征表示和分布概率计算。文本特征表示通过映射矩阵 M 将输入序列中的每个词映射为 n 维的特征向量，$M \in R^{v*n}$，再将序列中每个词对应的特征向量连接，形成一个大小为（N-1）*m 的特征矩阵 C。

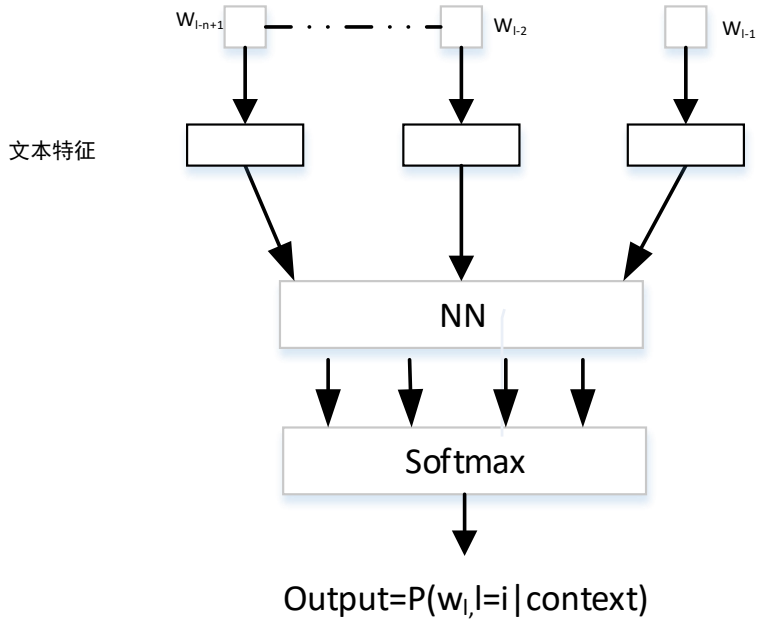


图 8-12 NNLM

NNLM 使用深度神经网络以特征矩阵为输入，映射输出一个概率分布向量 y，$y \in R^V$,y 的第 i 维表示输出为词典中第 i 个词的概率。

8.2.5 低资源语音识别系统方案

深度神经网络(DNN)模型的发展使得汉语、英语等大语种语音识别系统取得很高的识别率并已经走进市场、工业界。与此同时，也给不少资源匮乏语言的语音信号及文本信息处理带来了新的发展机遇。

在国家的大力支持下，少数民族语言，尤其是属于低资源语言的维吾尔语言语音识别研究有了一定的研究基础。神经网络模型在更小的语料资源上的高效的学习能力给资源匮乏语言的信息处理研究带来了新的机遇。有学者用基于卷积神经网络（CNN）声学模型与基于 N-gram 的统计语言模型（3-gram）结合进行维吾尔语语音识别；随后基于深度神经网络（DNN）和 HMM 结合的 DNN-HMM 声学模型与 N-gram 的统计语言模型(3-gram)相结合进行维吾尔语语音识别，虽然这些维吾尔语语音识别研究中用不同的声学模型，但针对语言模型都利用 N-gram 语言模型来进行语音识别。

为了确保中文语言识别系统能够发挥出更好的语音识别精确性能和实时性能。在本文中，结合语音识别基元层面进行了合理方案的选定。通过对系列声学模型的比对分析，进一步论述了传统的GMM-HMM 方案、实时性强的 DNN-HMM 模型和不同深度的CNN-HMM 模型的差异，语言模型方面，将 N-gram 语言模型和 RNN 语言模型展开比对。通过客观比对以及结果的剖析，发现后者能够进行大规模数据的高效处理，且处理结果的准确性效应更好。因此在声学模型方面，本文初步选定了 CNN-HMM 声学模型，语言模型方面初步确定选取 RNN 语言模型。

8.2.6 维-哈-柯低资源语言语音及文本预处理

随着神经网络模型的发展，低资源语言信息处理研究有了更好的发展。神经网络能够学习多个语言，并能将学习的内容迁移到新的环境中。这给很多低资源语言的信息处理研究带来了新的机会。但是，低资源语言的文本及语音信息预处理工作相对滞后，语料的质量和标准化工作跟不上等问题给这些语言的信息处理工作的发展带来困难。为了将少数民族语言文字信息处理中繁琐的预处理和资源整理工作简化、高效化、及梳理相关的工作，使低资源语言自然语言处理研究工作不必受制于文本信息的预处理工作。需要尽可能将预处理过程集成化、简化，用户界面统一化，让广大学者和用户不必学习该语言也能从事相关的信息处理工作。

维吾尔语、哈萨克语、和柯尔克孜语（以下称维-哈-柯语）等黏着性的少数民族语言具有相似的性质和特点，句中的词是自然分开的，构词和形态都是通过词干（或词根）后面连接不同词缀来派生出来的。其中词干是具有独立语义的单元且开放集，而词缀是辅助功能单元且闭集合。词缀功能强大，有构词词缀和构型词缀两种，构词词缀是改变词根的词义派生出新的词干，构型词缀决定一个词（词干）在句子中的语法作用。由于这种派生特性，这类语言在词素上有多种组合、

词和词性变化比较复杂，大大增加了词汇量。因此基于语素这样较小粒度单元的建模可以提供更强大的语义信息及更好的覆盖率，从而能建立更可靠的模型。为了就以上问题加以高效解决，在进行识别时，只对局部词干以及词缀规则展开分析，难以满足实际需要，语言中的每个单元的作用受到上下文信息的影响，所以还必须考虑句子上下文信息。

以往的研究是在单个语言上进行，而且没有构建一个完整的句子层面上的分析过程。曾经学者认为可以借助对词干库的建立将词缀库规则和统计完成有效结合，促使最终构建的方法，能够顺利获取到维吾尔语词干；还要部分学者尝试融合词性特征和上下文词干信息的维吾尔语词干提取模型。但目前为止，针对这些语言的词素提取方法及研究都是基于简单的词素形态分析和规则的方法为主，忽略了句子层面的上下文信息，这使得词素提取出现严重的歧义和不确定性现象。而且这些方法是零散的单个语言工具，缺乏通用性，难以推广应用。

为了就以上问题加以高效解决，在进行识别时，只对局部词干以及词缀规则展开分析，难以满足实际需要，语言中的每个单元的作用受到上下文信息的影响，所以还必须考虑句子上下文信息。因此，本章针对维-哈-柯等语言的已有的工作的基础上，在句子层面训练统计模型，并提供模型嵌入机制。各族研究者训练自己的统计模型并进行测试及比较。同时，可以将词嵌入向量法和聚类分析引入到词素切分及分析、词干提取等过程中，提高各类应用效率。

为了提高集成软件环境的通用性，尽量减少语言相关的人工工作量，我们的思路是将规则的收集、学习、训练、及切分等工作凝练成一个统一的软件框架，学习和训练部分与语言无关。每个语言只提供若干句子的词-词素平行语料库，和词缀库（闭集合）。通用软件直接从平行语料库中自动学习规则、语音和谐、上下文等信息，并对测试句子提供所有可能的切分序列结果。然后由独立的统计模型对该词素

序列进行排序。

8.2.6.1 低资源语言词素切分集成环境框架

该软件大概由 3 个层次构成。最底层是语音分析，由音素级的归一化、音节分析、拼写检查、及发音辞典生成等功能构成；中间层是低资源语言形态分析层，该层将词单元切分成所有可能的词干和词缀单元；最高层融合统计模型和中间层，从各种切分方式中选取最好的结果输出。如图 8-13 是本集成工具的系统框架。我们的目的是尽量减少语言相关的部分，纳入通用模版，减少手工工作。首先需要我们手工完成维-哈-柯三种语言的词-词素平行语料库。

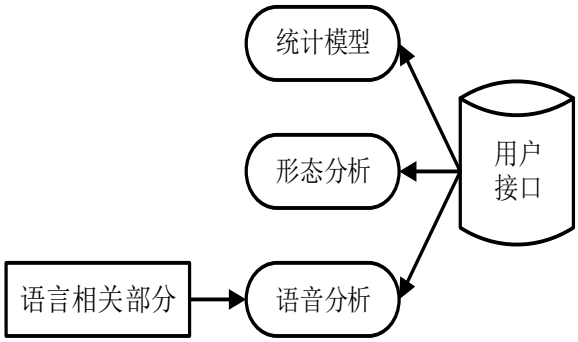


图 8-13 低资源语言词素切分框架

8.2.6.2 低资源语言的语音及文本特点

少数民族语言通常由于多源文化、方言、历史等原因对文本表达形式有着很高的不确定性和不规范性，或噪音。因此，需要首先从文字或音素层面上进行归一化以及标准化检测。我们首先将所有文本从各种编码形式规范化成统一的罗马字母编码形式（ASCII 码基本区），以便语言无关地统一处理。这些语言中通常音素和字母一一对应，因此口语和方言直接体现在文本当中，并产生个性化的拼写问题。在网络化时代标准化工作赶不上个性化网络语言的出现速度。如表 8-1 所示，一些字母由多种 unicode 代码来表达，需要对其进行归一化处理。

表 8-1 低资源语言字母的归一化表

统一的罗马字母	IPA	维吾尔语	哈萨克语	柯尔克孜语
A	ä	ە	ا	ە
		1749	1653	1749,1577,1607
e	ë	ې	ى	ئ
		1744	1609	1574,1569
y	j	ي	ي	ي
		1610	1610	1610
G	ɣ	غ	ع	ع
		1594	1593,65228,65227	1593,1594

对于噪音特性，较低成本的拼写检查方式是基于规则的分析方法。三个语言都有较清晰的音节结构，因此依据这些语言固定的音节规则，可以根据音节模版进行音节和规则层面上的拼写检查，能够检查出绝大部分拼写错误。

维-哈-柯语言中词干和词缀之间没有间隔标记，而且词缀表现形式多样，往往会出现复合后缀跟词干连接的形式。后缀决定和改变词干（词根）的语义和语法功能。词素由少量前缀加词干加词缀构成：prefix（前缀）+ stem（词干）+ suffix1+suffix2+...+suffixn （单后缀/复合后缀），如表 8-2 所示，其中词干是必要项。这类语言有少量外来前缀（有 6～7 个），且形态固定。

表 8-2 低资源语言词素结构

语言	词干（汉语）	词干+单后缀	词干+复合后缀
维吾尔语	maN(走)	maN+dim(我走了)	maN+al+may+mAn (我不能走)
	kAl(来)	kAl+di（来了）	kAl+diN+lar+mu (你们来了吗)
哈萨克语	jur(走)	jur+de(走了)	jur+al+dem (我能走了)
	al(拿)	al+de(拿了)	Al+deN+dar+ma (你们拿了吗)

柯尔克孜语	jol(路)	jol+da(在路上)	jol+dox+lar (同志们)
	oqu(学)	oqu+cu(学生)	oqu+cu+lar+din (学生们的)

词素切分中一个特殊的问题是语音的和谐规律，即词素的形态根据前后链接的音素产生形态变化或产生语音和谐。由于这些语言的文本和语音完全对应，语音的和谐直接体现在文本上。这个会导致歧义等问题，往往需要句子层面的语境分析。以往的研究都是根据规则总结变化形式，本章的方法是去掉人工工作，直接从词-词素平行语料中学习和训练语音的变化形式。

集成工具为语音识别与合成等研究提供发音词典生成器，能够建立音素、音节、词素、词等多个层面上的发音词典和文本。在音节层面上根据音节模版检测拼写错误。该方法简单且高效、适合低资源语言，其流程在如图 8-14 所示。

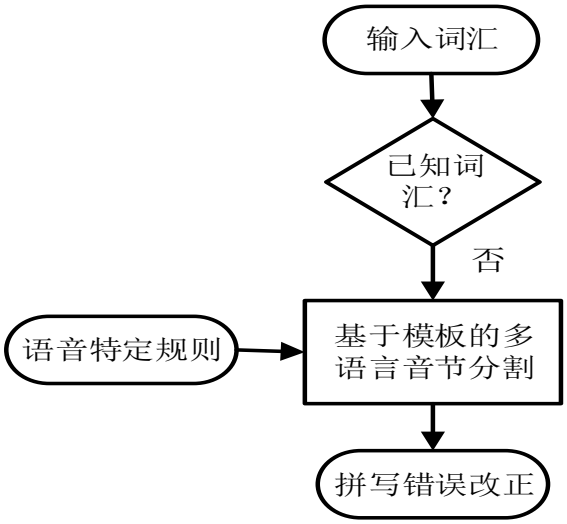


图 8-14 低资源语言拼写检测流程

8.2.6.3 低资源语言的形态分析

低资源语言形态分析根据每种语言提供的词-词素平行语料进行各个词素单元和链接形式的学习。词-词素平行语料库如表 8-3 所示。

表 8-3 低资源语言词-词素平行语料示例

维吾尔语：
silAr xinjaNniN viqtisadiy tArAqqiyatiGa nahayiti zor tOhpA qoxtiNlar.
silAr xinjaN+niN viqtisad+iy tArAqqiyat+i+Ga nahayiti zor tOhpA qox+tiN+lar.
哈萨克语：
sEndEr xenjyaNneN Ekonomeyka damwena otE zor ENbEk korsEtteNdEr.
sEn+dEr xenjyaN+neN Ekonom+ey+ka damw+e+na otE zor ENbEk kor+sEt+teN+dEr.
柯尔克孜语：
svzdAr xvnjaNdin Akonomikasinin UnOgOsOnU zoor tUkpUU qoxtuNar.
Svz+dAr xvnjaN+din Akonomika+si+nin UnOgOsO+nU zoor tUkpUU qox+tu+Nar.
句意：你们为新疆的经济发展作出了巨大贡献。

为了减少或避免人工语料出现的错误，将闭集合，即 124 种词缀附加成分的各种表现形式单独收集并对其在各个语言中进行标准化。一方面提高系统的检错和纠错能力，能给人工语料进行检错。另一方面完成了附加成分（对于黏着性语言主要是词缀）的全面分类和标准化工作。每个附加成分严格根据语义和语法功能分解，并需要收集相应的语音和谐引起的表现形式。如：词缀 gha, qa, gA, kA, ikA, igha, iqa…等。

词素边界上的音素根据语音和谐规则改变其表面形式。当发音准确地表达时，可以在文本中清楚地观察到语音和谐。低资源语言切分模块根据词-词素平行语料库学习所有的词素序列以及语音和谐的各种变化形式。由于词缀之间的形态变化基本已收集，语音和谐的形态重点学习词干与第一个词缀的边界。以往的工作采用的是根据规则来判定语音和谐的形态，该方法对于低资源语言处理不适用且很难收集所有形态规则。该工具不仅减少不确定性，而且提高了可靠性，并很大程度地改进了黏着语言文本处理效果。

为了实现功能上的可扩展性，词素序列的学习和语音和谐的形态学习保障通用性原则。该工具通过模版搜索算法匹配词干库和词缀库，

将大规模文本语料库中的每个候选词切分成所有可能的词素序列形式，并把这些词素送入一个独立的统计模型中，从候选词素序列中排序出前 N-best 个最佳序列。词素切分流程如图 8-15 所示。从词-词素训练语料中获得的词素序列为独立的统计模型提供训练语料。

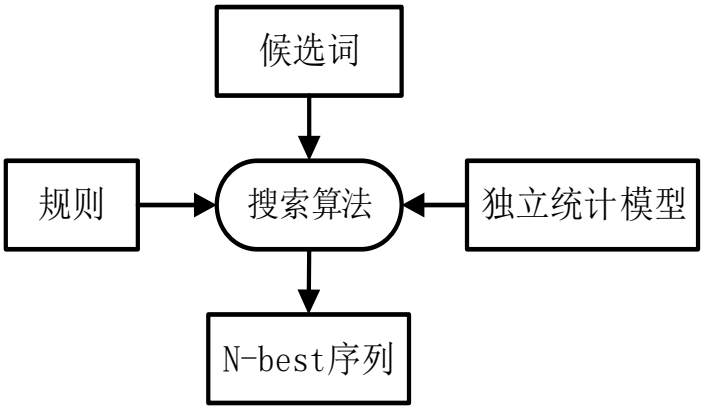


图 8-15 低资源语言词素切分流程

8.2.7 语音识别评价标准

评价语音识别系统性能的好坏通常需要指定统一标准的评价指标。虽然不完美，但目前比较公平的评价指标有，音素错误率(PER)、字错误率(CER)、词错误率(WER)和句错误率(SER)。本文主要用 WER 来对所建立的系统进行评价。该指标定义为如下，假设 N 为测试集中总词数目，解码后插入的词数为 I ,删除的词数为 D 以及被替换的词数为 S，那么词错误率可以用以下公式计算：

$$\mathrm{WER} = 100 * \frac{\mathrm{I} + \mathrm{D} + \mathrm{S}}{\mathrm{N}} \%$$

准确率为：

$$\mathrm{Accuracy} = 100\% - \mathrm{WER}$$

8.3 语种识别技术

语言是沟通交流的方式之一，在人们的日常生活中有着无可替代的作用。语种识别，其目的是让计算机能够预测出语音的语种类别。作为语音识别的前置技术，语种识别为后续的语音识别奠定了良好的基础。在全球化日益加剧的进程中，多文化、多语言相互交融，语言不通导致的交流障碍日益凸显。为了解决语言沟通上的困难，所采取的措施就是对语言种类进行划分，因此，语种识别技术应运而生，成为了模式识别领域的一个研究热点。下面简要概述了语种识别技术的四个应用场景，进而体现了其研究的必要性和研究意义：

1. 复杂条件下的语音识别

目前的语音识别系统只能完成单一语种的语音识别，而无法实现多语言的语音识别。因此语音识别系统在对待测语音片段做出处理时，需要设置一个语种识别子系统来确定输入语音片段的语种类别。例如，微信中的语音转文字，只支持用户使用中文功能；苹果的语音助手 siri，最初只适用于英语用户，而其他的用户无法使用此功能。科技的发展，导致不同用户的需求激增。若首先利用语种识别技术对各语言进行分类，那么语言不通的问题就可以被轻松解决。

2. 多语言辨识系统

在国际性场景，需要对同一段语音内容以不同的语种进行表达。例如，世博会、奥运会、国际会议等，都需要配备相应的跨语言话务员来进行同一语音内容的不同语种传输。如果可以在一系列的国际性活动中引入语种识别技术，首先对不同语言种类进行划分识别，然后对接不同的话务员，这将会大幅提高工作效率。

3. 语言学的发展

在一些少数民族地区，少数民族语言鲜为外人所知，也出现了消退的趋势，这就给语言学的研究带来了挑战。因此，语种识别技术可以解决这一难题。面对一段未知语种的语音，可以先利用语种识别技

术检测该语音的语言种类，然后翻阅文献进行相关研究。

4. 国家安全

在刑侦领域，语音证据也很重要。例如，利用语种识别可以对嫌疑语音记录等进行分析处理，进而判别出嫌疑人的身份信息等。除此之外，涉及到国家安全的语音情报工作，也需要语种识别技术的辅助。它可以针对性地检索出关键的信息，滤除掉无用信息，减少人工检查的时间和成本，提升效率。

语种识别技术，至今已有五十多年的历史，也取得了不错的成就。其中，确定条件下的语音识别，已经达到了人工标注的效果。大数据和人工智能时代的到来，使得语音更复杂，用户需求更多样化。因此，语种识别开始走向混淆语言的识别，如方言，少数民族语言和小语种等。尽管语种识别发展良好，但仍然面临很多问题，例如鲁棒性、语音时长等因素的影响，这使得其发展空间有待提升。

8.3.1 语种识别研究背景

语种识别技术，可以追溯到 20 世纪的 70 年代，早期由于数据库等多种因素的原因，语种识别技术进展非常缓慢；由于数据库不统一以及缺乏公开的数据库，研究结论很难进行比较和分析，从而使得语种识别停滞不前。1992 年，俄勒冈科学技术院公布了电话语音的语料库（Oregon Graduate Institute of Science and Technology Multi-language Telephone Speech Corpus，OGI-TS）。之后，OGI-TS 语料库逐渐成为了语种识别领域的统一数据库，研究者们也开始在学术界和工业界开展了更广泛的研究。1996 年，国家标准技术研究所（National Institute of Standards and Technology，NIST）发布测评竞赛（Language Recognition Evaluation，LRE），该比赛也成为了国际上颇具影响力和权威性的比赛，同时助推语种识别技术从纯粹的理论研究开始走向应用，语种识别的实用性也不断增强。语种识别的方法主要有：基于音素特征的语种识别、基于底层特征的语种识别和基于深度

学习的语种识别。

8.3.1.1 基于音素特征的语种识别

基于音素特征的语种识别，是通过不同语言的音素的不同搭配来实现的。其整体思路是，首先得到一个音素识别器，之后通过不同语种的音素组合，来建立不同语种的语言模型，进而实现语种识别。音素识别器结合语言模型（Phone Recognizer followed by Language Model，PRLM），该方法首先获取语音的音素的序列，然后提取 N-gram 特征，最后由统计特性建立语言模型(Language Model，LM)）。除此之外，研究者们在 PRLM 语种识别系统的基础上，并行构建了多套 PRLM 系统，来实现语种识别，该方法是并行的音素识别器（Pallel Phone Recognizer，PPR），并产生了语言模型—PPRLM（Parallel PRLM）。支持向量机 SVM 的出现，提供了新思路：音素识别器与支持向量机的结合，取得了较好的识别效果。近些年来，全差异建模方法的引入，也取得了良好的研究成果。后来，研究者们提出了基于 i-vector（identity-vector）的语种识别方法，该方法不仅可以降低音素特征的维度，还可以降低语言模型的复杂度，因此在语种识别研究上表现出了更佳的的性能。

8.3.1.2 基于底层声学特征的语种识别

基于底层声学特征的语种识别，是利用底层声学特征所能够描述的声学单元的统计特性差异来对语种进行分类。研究表明，声学特征有：线性预测倒谱系数，梅尔频率倒谱系数，感知线性预测系数、移位差分倒谱特征等。Wang Maorong 等人提出了将 SDC 特征与 GMM-UBM 结合的方法。其中，SDC 特征是由 MFCC 特征的差分扩展得到的，可表示长时的声学单元；而 GMM-UBM 可表现声学特征的分布。后来研究发现，说话人、信道、噪声等会影响语种识别的性能，研究者们又提出了几种方法，分别为倒谱域减均值、声道长度规整，来尽可能消除非语种信息的影响。

研究者们在对底层声学特征进行不断改进的同时，引入了新的方法。由于 GMM-UBM 模型无法对易混淆语种进行分类判决，因此，在GMM模型的基础上，研究者们又提出了3种方法：一是GMM-MMI方法，该方法是利用最大互信息准则 MMI 来训练 GMM 模型；二是GSV-SVM（GMM Super Vector-SVM）方法，该方法是利用支持向量机模型，对 GMM 的超矢量进行建模；三是 Model Pushing 方法，该方法利用方法二，得出语种的 GMM 模型。除此之外，在联合因子分析方法的启迪下，Deha 等人运用了因子分析，来消除信道噪声的干扰，提升分类模型的性能，进而得到了 i-vector 的方法。后来，研究者们提出了全差异空间建模的方法，该方法也成为了主流系统之一。

8.3.1.3 基于深度学习的语种识别

在 90 年代，研究者们开始使用神经网络。但是由于种种限制，未能顺利进行。2009 年，Montavon 等人用神经网络提取特征，这是神经网络首次应用成功。2014 年，Ignacio Lopez-Moreno 等人提出DNN 网络，获得的效果优于 ivector 方法。同年，Lei 等人提出了 CNN 网络，表现出了很好的效果。2015 年，Richardson 等人利用 DNN 网络提取瓶颈层特征 BNF，取得了更佳的性能。

2014 年，蒋兵等人在 DBF / i-vector 系统上，利用最大互信息熵 MMI 对 DBF 特征进行调整。Daniel 等人提出 TDNN，解决了 DNN 网络只能在短时语音上建模的问题。Gonzalez 等人利用 LSTM-RNN，解决了传统 RNN 网络的问题，即梯度消失、梯度爆炸等问题。Geng 等人利用注意力机制模型，重点关注一段语音中与语种信息相关的部分。可以发现，基于端对端神经网络的语种识别系统目前仍处于探索阶段，是深度学习在语种识别上的一大进展，研究者们正在针对该方法展开更深入的研究。

8.3.2 语种识别原理

对于计算机来说，语种识别是典型的分类识别问题，它包括特征

提取、模型训练、测试识别等阶段，其系统架构如图 8-16 所示。在训练过程，首先进行特征提取，该特征能够反映语种的区分性；然后根据相应的算法，训练出每个语种对应的模型，并进行保存。在识别过程，对于待测的语音信号，首先提取特征，并送入之前训练阶段产生的模型中进行匹配，最后根据相似度，判断待测语音的语言种类。其中，特征提取、模型的分类判别部分能够对语种识别的性能产生较大的影响。因此，要提高识别性能，就需要选择合理的特征提取和分类判别模块。

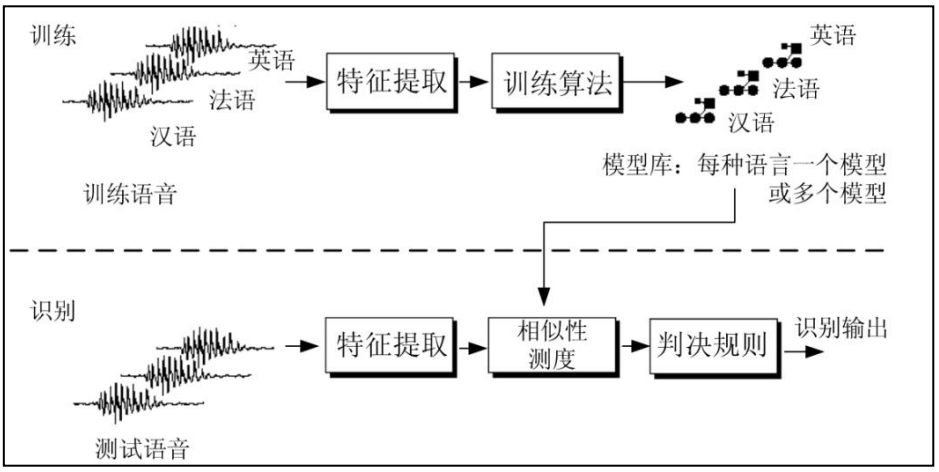


图 8-16 语种识别系统框图

8.3.2.1 特征提取

语音信号复杂多变，只有在很小的间隔内（15-30ms）是平稳的，因此要对语音信号进行短时分析。除此之外，语音信号中还包含噪音、情感、说话人等多种信息，会影响语种识别的性能。因此，在进行语种识别时，要先进行预处理，如去噪、消除静音等。

语音信号的特征包括底层声学特征、韵律特征、词法特征和语法特征。其中，底层声学特征是最基础的特征，其它特征都是在此基础上产生的。越高的特征，语种的分类效果越佳；越低的特征，冗余的信息多，导致语种判别效果降低，在特征提取上的代价就越低。而对

于底层声学特征来说，其提取过程简单易实现，没有词法和语法特征提取繁琐。

语种识别技术中，基于底层特征的方法已经远超其他的特征。下面将介绍语种识别中常用的底层声学特征，有线性预测倒谱、梅尔频率倒谱和 Fbank、移位差分倒谱和语谱图等。

8.3.2.2 线性预测倒谱

线性预测倒谱（Linear Predictive Cepstral Coefficient，LPCC）是一种最早应用于语音分析的参数，它的优点是：去掉了语音的激励信息，反映了对声道的回应，计算量较小。在一段语音中，可以利用最小均方误差，得到一系列的预测系数，就是 LPC 参数。LPC 参数对应的声道模型可以表示如下：

$$H(z)=\frac{s(z)}{E(z)}=\frac{G}{1-\sum_{i-1}^{P}a_i z^{-i}}=\frac{G}{A(z)}$$

式中，G 为增益常数，$a_i(i-1,2,\ldots,p)$是 LPC 参数，P 是模型的阶数。那么，对于输入μ_n，输出 S(n)可表示为：

$$s(n)=\sum_{i=1}^{P}a_i\,\mu(n-i)$$

LPCC 参数很好地反映了语音信号的声道信息，它与 FFT 参数相比，在描述短时谱上更加细致，但是其在所有频率上都采用线性关系来逼近语音信号，并不符合人耳的听觉特性；而且它假定信号存在一种线性预测的结构，这对于清音来说是强加了一种错误的表示结构，这些都会影响系统的性能。研究者们为了改善上述问题而提出 MFCC 参数和 FBank 参数。

8.3.2.3 梅尔频率倒谱

20 世纪 80 年代，梅尔频率倒谱（MFCC）被提出。之后，MFCC 的应用开始变得广泛，如说话人识别、语种识别等。MFCC 基于听觉系统，在 1000Hz 以下，人耳对频率的感知是有规律的；超过 1000Hz，人对声音的感知是不规律的，这就导致人对低频的信号更为敏感。鉴

于人耳的听觉特性，MFCC 则是在 Mel 频率上产生的，它描述了一种非线性的关系，可以用以下公式近似表示，其中，f 是频率，图 8-17 展示了频率由线性到非线性的转换。

$$Mel(f) = 2595 \times log(1 + \frac{f}{700})$$

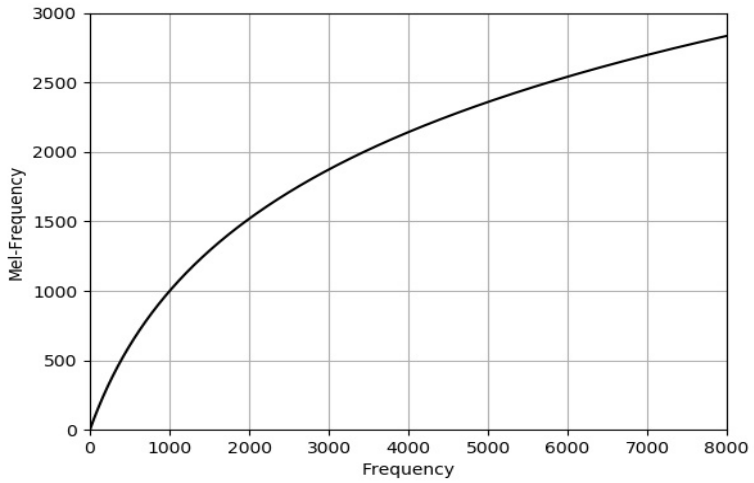


图 8-17 Mel 频率图

MFCC 特征利用了人耳的听觉特点，对语音信号没有假定和限制。因此，它数比 LPCC 参能更好，更能产生较好的效果。MFCC 的计算过程如图 8-18 所示。

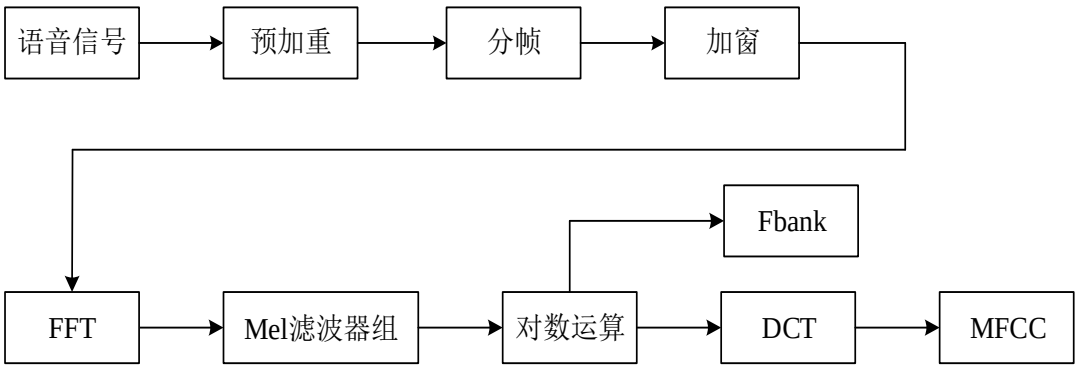


图 8-18 MFCC 的计算过程

（1）预加重：让输入信号通过高通滤波器，滤波器的函数：

$$H(z) = 1 - \alpha z^{-1}$$

其中，α是 0.9-1.0 范围内的数值，取 0.97 效果最好。预加重的目的是：提高语音信号中频率高的成分，平坦化语音信号。

（2）分帧、加窗的处理：语音信号进行短时分析时，需要分帧操作。相邻的语音帧之间一般会存在重叠交叉部分，被称为帧。通常语音信号的帧长为 25ms，帧移为 10ms。分帧完成后，用窗函数（如 Hamming 窗）滑过语音帧，可以连接语音帧，使其连续，过渡平滑。

（3）快速傅里叶变换（Fast Fourier Transform，FFT）：经过步骤（2）的操作后，对语音进行 FFT 操作，可以得到语音信号的频谱；之后对其进行取模的平方，可以得到幅度谱。

（4）Mel 滤波：将步骤（3）的输出通过一组 Mel 滤波器组，其包含的滤波器个数为 M，中心频率是 f(m)。m的取值为 $1,2,3\ldots,M$，M 为 22-26 内的数值。f(m)之间的距离是不同的，与 m 值同步变化，并且满足公式：

$$Mel(f(m)) - Mel(f(m-1)) = Mel(f(m+1)) - Mel(f(m))$$

Mel 滤波器的频率响应定义为：

$$H_m(k) = \begin{cases} 0 & ,k < f(m-1) \\ \dfrac{2[k - f(m-1)]}{[f(m+1) - f(m-1)][f(m) - f(m-1)]} & ,f(m-1) \le k \le f(m) \\ \dfrac{2[f(m+1) - k]}{[f(m+1) - f(m-1)][f(m) - f(m-1)]} & ,f(m) \le k \le f(m+1) \\ 0 & ,k \ge f(m+1) \end{cases}$$

式子中，$\sum_{M=0}^{M-1} H_m(k) = 1$。

Mel 滤波器的作用：一是使频谱变得平滑，减小谐波对语音的扰动作用；二是降低计算量。因此，基于 MFCC 特征的语种识别，不会受语音音调的影响。

（5）Fbank 特征（Filter banks）：经过步骤（4），得到 Mel 滤波器组的输出，并取对数能量，之后对能量谱进行 Mel 转换，就得到了 Fbank 特征，计算公式如下：

$$s(m) = ln(\sum_{M=0}^{M-1}|X(k)^2|\,H_m(k)),0 \le m \le M$$

式中，$s(m)$是对数能量，$X(k)$是输入信号的 FFT 变换，M 是 Mel 滤波器的个数。

（6）离散余弦变换（Discrete Cosine Transform，DCT）：通过步骤（5）得到对数能量，进行 DCT 变换，可以得到 Mel 系数，为 L 阶。L 为 12-16，M 指 Mel 滤波器的数目，计算公式如下：

$$C(n) = \sum_{m=0}^{N-1} s(m)\, cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1,2,\ldots,L$$

（7）动态差分（一阶差分、二阶差分）：MFCC 是静态参数，无法全面反映性能。研究表明，将 MFCC、其一阶差分、二阶差分结合，能够很好的反映其识别性能。一阶差分、二阶差分的计算公式为：

$$d_t = \begin{cases} C_{t+1} - C_t & ,t < K \\ \frac{\sum_{k=1}^{K} k(C_{t+k} - C_{t-k})}{\sqrt{2\sum_{k=1}^{K} k^2}} & ,其他 \\ C_t - C_{t-1} & ,t \ge L - K \end{cases}$$

其中，d_t是一阶差分；C_t是倒谱系数；L 是阶数；K 是时间间隔，K= 1 或 2。得到一阶差分后，通过以上公式，则可得到二阶差分。

8.3.2.4 移位差分倒谱

移位差分倒谱特征（Shifted Delta Cepstra，SDC），通过一阶差分谱生成，它有四个参数（N，d，P，k），其中 N 是静态参数的维数，d 是帧之间的距离，P 是跳帧的长度，k 是帧的个数，SDC 特征示例如图 2-4 所示。假定$c(t) = [c_0(t), c_1(t),\ldots, c_{N-1}(t)]^T$。则差分公式为：

$$\Delta c_i(t,j) = c_i(t + jP + d) - c_i(t + jP - d),0 \le i \le N - 1,0 \le j \le k - 1$$

其中，d 为 d 阶差分，P 为跳跃间隔，k 为跳跃次数。由此，可以得到 k 维的差分值：$\Delta c_i(t,i) = [\Delta c_i(t,0), \Delta c_i(t,1),\ldots, \Delta c_i(t,k)]$，则第 k 帧的新 SDC 为：将 c(t)的各维通过上述公式计算得到的 kxN 个差分值连接起来：

$$SDC(t) = [\Delta c(t,0), \Delta c(t,1),\ldots, \Delta c(t,N-1)]$$

综合以上公式，不难看出，第 t 帧、第 t+P 帧的 SDC 特征存在重合，重合维度为(k-1)xP 维。这样的优势有两点：一是 SDC 特征包含了更多的时序信息；二是 SDC 特征是连续的。在 SDC 特征中，要加入原始静态特征 c(t)，即

$$SDC(t) = [c(t), \Delta c(t,0), \Delta c(t,1), \ldots, \Delta c(t,N-1)]$$

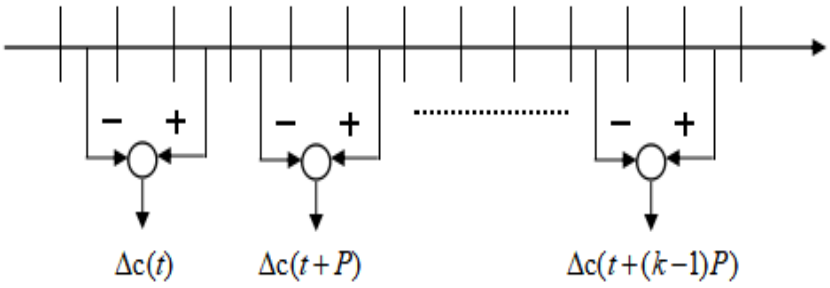


图 8-19 SDC 特征示例

8.3.2.5 语谱图

语谱图，是语音信号的一种图像表示，其横轴是时间，纵轴是频率，坐标是语音信号能量。语谱图中的颜色深度反映了能量值，颜色越深，语音的能量值就越大。除此之外，语谱图中也包含基频、共振峰参数，它们既可以表现出时域波形的特点，也可以反映频谱图的特点。语谱图示例，如图 8-20 所示。相比较于传统的特征来说，语谱图是语音信号比较好的一种特征，因为它对于原始语音数据的加工处理是最少的，并且它能够保留较多的原始语音信息，因此具有信息无损的特性。

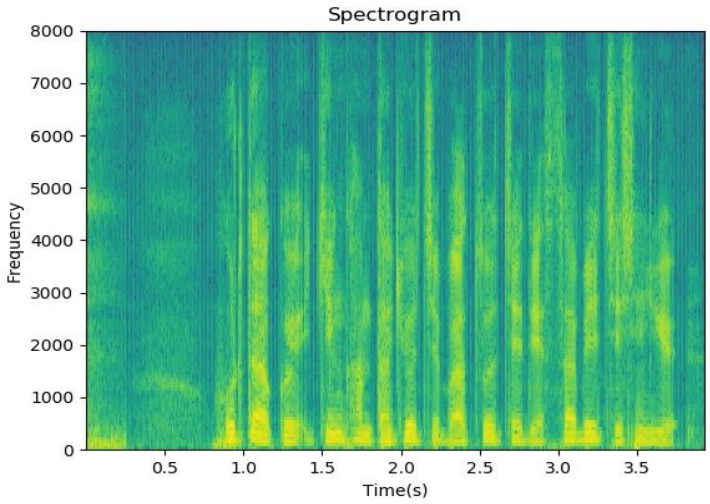


图 8-20 语谱图示例

从语谱图中可以看出，其中有横线、竖直的线条、杂乱的花纹等，它们各自有不同的表示含义。其中，横线是共振峰，竖直的线条是语音信号中一个基音，乱纹的深浅程度表示噪声能量的分布。在一段语音音频中，语谱图代表了语音信号的最原始的信息，因此可利用图像处理的方法对语谱图进行相关的研究。

8.3.3 声学模型

声学模型能够反映出声学特征所包含的语音信息，在语种识别中较为重要。目前主流的声学模型为：概率模型、可区分性模型。概率模型，是通过计算特征的概率来实现的。不同的语言有不同的特征分布，相似语种的特征分布有相似的部分，而非相似语种的特征大不相同。因此，根据这一特性，可以用概率密度的大小，代表语种的声学特征与模型的相似程度。可区分性模型描述了在高维空间中的语种的特征分布情况。对于不相同的语种，其声学特征分布较远，分类简单；而对于相同的语种，其声学特征分布较近，难于区分。因此，可以用置信度计算，其范围为[0，1]或[-1，1]。

8.3.3.1 GMM 模型

对于语音数据来说，其声学特征分布复杂，无法用简单的模型表示。为了逼近其分布特性，研究者们提出了高斯混合模型（Gaussian

Mixture Model，GMM）。其基本思想是：首先提取语种的声学特征，然后利用高斯函数，得到语种的特征分布。则 GMM 分布模型可以用公式表示为：

$$p(x|\Lambda) = \sum_{m=1}^{M} w_m \frac{1}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}(x-\mu_m)^T \sum_m^{-1}(x-\mu_m)\right\}$$

其中，x是特征向量，M 是高斯分量；$\mu_m, \sum m$表示均值、方差；w_m表示权重，且满足$\sum_{m=1}^{M} w_m = 1, w_m \geq 0$。GMM 模型的结构如图 8-21 所示。

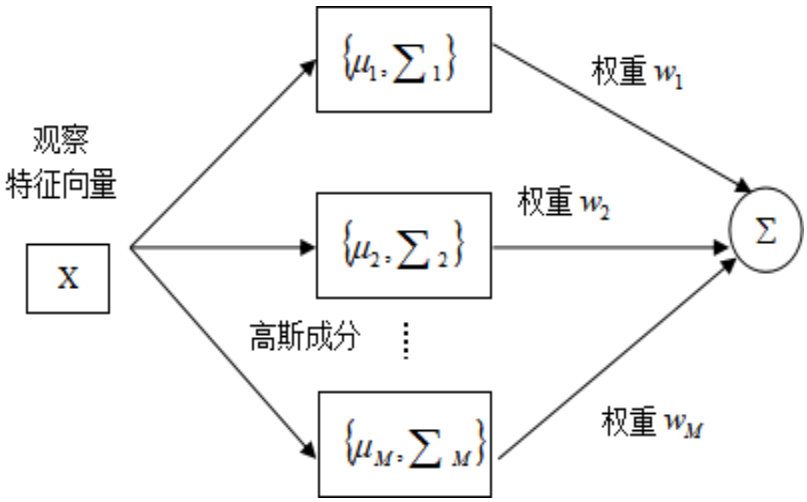


图 8-21 GMM 模型结构

GMM 模型能够表示语种的特征分布，但是需要充足的训练样本，而数据量有限时，则会出现过拟合现象，使模型参数估计不准确。为了解决该问题，研究者们提出了通用背景模型方法（Universal Background Model，UBM）。GMM-UBM 模型，首先训练出一个通用背景模型Λ，然后得到每种语种的 GMM 模型Λ_l，则待测语音信号对目标语言的似然度使用Λ_l和Λ两种模型计算。后来，研究者们提出了最大互信息准则（Maximum Mutual Information，MMI）方法。假设有 N 个语音样本，即$S = \{s_1, s_2, \ldots, s_N\}$，每个语音样本的特征向量为$X = \{X_1, X_2, \ldots, X_N\}$，其中$X_n = \{x_{n,1}, x_{n,2}, \ldots, x_{n,3}\}$，则 MMI 准则的目

标函数可表示为：

$$L_{MMI} = (\Lambda_l|X) = \frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n|\Lambda_{l_n})p(l_n)}{\sum_l p(X_n|\Lambda_l)p(l)} = \frac{1}{N}\sum_{n=1}^{N}p(\Lambda_{l_n}|X_n)$$

其中，l_n是语种类别，$p(\Lambda_{l_n}|X_n)$是语音是某一类别的后验概率，X 是输入信号，N 为样本个数。

8.3.3.2 SVM 模型

支持向量机用于统计学习，它是通过统计学方法，通过使不同样本的间隔最大，进而获得分类超平面。当样本是线性、可分类时，找到分类超平面，使样本之间的间隔最大，可得到线性的分类器，即为线性的可分支持向量机；当样本是非线性，无法分类时，则要使用核函数，来得到非线性的支持向量机。下面将具体介绍支持向量机的分类原理。

对于二分类的问题，假设数据集$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，其中，特征向量$x_i \in X \in R^n$，样本的标签$y_i = \{-1, +1\}, i = 1,2,\ldots,N$，N 是样本数。通过公式，可以找到支持向量机的最优分类平面。

$$\min_{w,b}\frac{1}{2}\|w\|^2$$
$$s.t.\ \ y_i(w \cdot x_i + b) \geq 1, i = 1,2,\ldots,N$$

其中，w，b 是参数。假设w^*，b^*是最佳解，则分类平面为$w^{*T} + b^* = 0$，如图 8-22 所示。其中，圆圈为正样本，“十”字为负样本，黑色实线为最优超平面，黑色虚线上的样本为支持向量。从图中可看出，支持向量机，不仅可划分正、负样本，还可使它们之间的距离最远，这可以使得支持向量机对未知的新样本有着很好的分类能力，分类决策函数为公式。

$$f(x) = sign(w^T x + b)$$

其中，w，b 是超平面的参数，w 为权重值，b 是偏置值，x 是输入样本，sign()为符号函数。

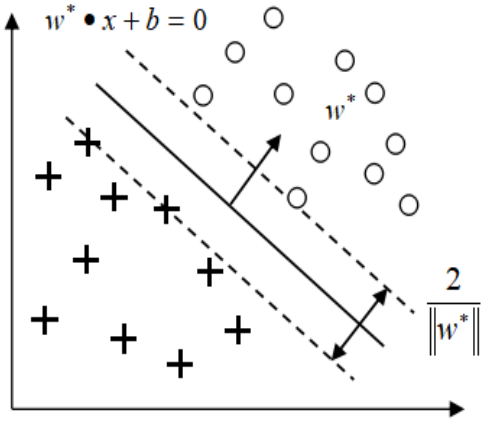


图 8-22 SVM 最优超平面

在实际应用中，数据是非线性的，因此需要非线性的支持向量机，其重点是运用了核函数。它的基本思想是，首先将非线性样本投射到高维空间中，使其具有线性规律，然后在高维空间中生成对应的支持向量机。另外，还可以通过以下公式，获得线性支持向量机问题的解：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i{}^T x_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \sum_{i=1}^{N}\alpha_i y_i = 0, \alpha_i \geq 0, i = 1,2,\ldots,N$$

假设式以下公式的最优解为$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$，则原始问题的最优解为：

$$w^* = \sum_{i=1}^{N}\alpha_i^* y_i x_i$$

$$b^* = y_j - \sum_{i=1}^{N}\alpha_i^* y_i (x_i y_i)$$

其中，y_j是样本标签。以下公式都用到了内积公式，因此无需知道原始特征在高维空间中的表达式，只需要知道其内积的表达式，也就是核函数，就可以进行求解。核函数有以下几种：

（1） 线性核函数：

$$K(x,z) = x^T z$$

（2）多项式核函数：

$$K(x,z)=(\gamma x^T z+r)^p,\gamma>0$$

（3）径向基核函数（Radial Basis Function，RBF）：

（4）Sigmoid 函数：

$$K(x,z)=tanh(\gamma x^T z+r),\gamma>0$$

其中，$K(x,z)$为核函数，γ，r，p分别为核函数的参数。

虽然 SVM 可以实现分类问题，但也存在不足之处。其不足是，它仅面向长度固定的向量，无法处理不同长度的向量。语音是可变序列，无法满足 SVM 的要求。

因此需要进行转换，使其满足 SVM 的函数形式。

8.4 文字识别技术

近年来，在国内汉语智能信息处理技术研究成果的先导作用下，维吾尔语文本信息的智能处理也得到了长足发展。本节主要以新疆大学研究团队近几年研究工作进展为主，介绍维吾尔文字母识别，单词分割，单词识别等方面的研究方法和主要对比研究工作。

8.4.1 联机手写维吾尔文字母识别

字母识别系统的结构如图 8-23 所示，主要由预处理、特征提取、聚类分析、建立字母模板库以及字符识别等模块构成。

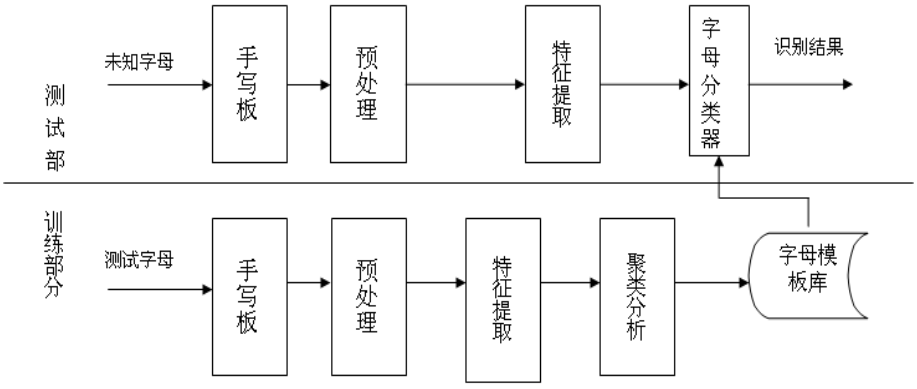


图 8-23 维吾尔文手写字母识别的总体结构

8.4.1.1 联机手写字母识别对比研究

特征提取是字符识别的核心步骤，直接影响到字符识别的正确率。用基于坐标归一化的特征提取(Normalization-cooperated feature extraction, NCFE)提取局部方向直方图特征，并用 Fisher 线性鉴别分析（FLDA）对特征进行降维，以减小计算复杂度。在分类阶段，分别采用修正的二次判决函数(MQDF)、具有判断学习型的二次判断函数(DLQDF)、学习矢量量化(LVQ)和具有径向基本函数内核的支持向量机分类器(SVC—rbf)等四种分类器。

在表 8-4 中比较了 MQDF、DLQDF、 LVQ 和 SVC-rbf 等 4 种分类器的性能。在各种分类器评估中，把伪二维矩归一化（P2DMN）方法和 8 方向 5*5 块的 NCFE 特征提取方法作为基准。原始特征经 FDA 变换后，从 200 维降到 120 维。采用 SVC-rbf 实现时，其核的参数变化正比于方差估计，且系数的上限设定为 l0，σ^2=0.3。试验结果表明，采用 DLQDF 分类器的情况下识别率取得了最高识别结果 88.93%。DLQDF 与 SVC-rbf 的识别结果差异不大。而采用 LVQ 分类器得到的识别结果比 DLQDF 要差 2 个百分点，这应该与分类器自身的分类性能有很大的关系。

表 8-4 不同分类器下的性能对比（%）

特征		MQDF	DLQDF	LVQ	SVC - rbf
NCFE	正确率	87.98	88.62	86.23	88.50
	10 候选	98.95	99.10	98.88	98.13
NCFE+ 几何特征	正确率	88.77	89.08	87.77	88.92
	10 候选	99.16	99.14	98.96	98.87

8.4.1.2 基于 DTW 的联机手写字母识别

DTW 匹配算法能在笔迹有一定变动的时候仍能达到良好的效果，

因此采用了时间弯折算法(Dynamic Time Warping)。

假设有两个特征序列 A, B，其中 A 为模板特征序列，B 为测试特征序列，

A=al,a2,...,ai,...,aI ,

B=bl,b2,...,bj,...,bJ

A, B 特征序列间相对应的时间变化关系可以由下面的时间规整函数表示:

f=c(1),c(2), ..., c(k), ... c(N)，

其中 c(k)=(I(k),J(k))，代表在作 k 次特征匹配时，测试特征序列 B 中第 J(k)帧与模板特征序列 A 中第 i(k)附帧比较，c(k)可视为是 i-j 平面上的一个点，它随者参数 k 在 i-j 平面上的移动形成一条曲线，称为“时间弯折匹配曲线”或“匹配路径”。

设 d(I(k),J(k))或 d(c(k))表示将模板中第 i 帧与测试序列中的第 i 帧进行匹配的局部匹配距离(我们采用欧氏距离)。当两个特征序列之间有相对瞬时时间变化时，为了最大限度地承认这种波动，DTW 的目标应该是找使得两特征序列的总体平均匹配距离满足公式:

$$D(A,B) = min\frac{\sum_{k=1}^{k} d\big(c(k)\big)w(k)}{\sum_{k=1}^{k} w(k)}$$

匹配路径 f，将其作为模板 A 和测试序列 B 的匹配路径，将这时的匹配距离作为测试序列 B 和模板 A 的匹配距离，其中 w(k)为匹配点 c(k)匹配距离的加权系数。

表 8-5 示出了四种不同类型的字母和平均识别率的识别率。此实验主要集中在封闭和开放测试的识别性能。基于开放测试不同类型的字母的候选率的实验结果示于表 8-6 中。我们获得了 70.73％的维吾尔文字母平均识别结果，识别率一般。但是五个候选之一的平均识别率是 95％，它是令人满意的。结果表明，所采用归一化，特征提取，动态聚类和分类方法，有效的解决了字母识别问题。通过识别后的再分析，得到相互混识严重的字母集。

表 8-5 不同类字母的识别率 （%）

	第一类	第二类	第三类	第四类	总平均识别率
训练集直接作为测试集	78.64	74.11	67.54	78.74	75.16
另外 25600 样本作为测试集	74.67	70.42	63.33	72.02	70.73

表 8-6 字母识别 5 个候选的统计率

字母识别候选	第一类	第二类	第三类	第四类
1	74.67	70.42	63.33	72.02
2	86.85	86.09	80.43	88.41
3	90.72	90.82	87.57	90.75
4	92.84	92.73	91.13	94.75
5	94.34	94.18	93.15	95.86

8.4.1.3 基于多分类器集成的手写字母识别

集成系统由$C_1, C_2, \cdots, C_n$等 n（n=5）个分类器组成，待识别维吾尔文字符 A 经分类器C_i识别后，产生 m 个候选字输出$C_{i1}, C_{i2}, \cdots, C_{im}$，每个候选字对应的相似度为$S_{i1}, S_{i2}, \cdots, S_{im}$并且有$1 \geq S_{i1} \geq S_{i2} \geq S_{i3} \geq \cdots \geq S_{im} \geq 0$。

各个分类器将各自的候选字及相似度输出到集成判决器中，集成判决器根据一定的策略进行处理，从而产生最终的识别结果。由此可见，集成判决器是多分类器设计的关键。在设计集成判决器时，定义个分类器的候选字C_{ij}的置信度函数为：

$$D(i,j) = \lambda_i p_j s_{ij} + \sum_{k \neq i} \lambda_k \, p_{n(k,i,j)} s_{kn(k,i,j)}$$

其中，各个分类器的加权因子分别为$\lambda_1, \lambda_2, \cdots, \lambda_n$其中有$0 \leq \lambda_i \leq 1$；每个候选字的加权因子为$p_1, p_2, \cdots, p_m$。

基于置信度最大原则的判决策略为：

$$D(M,N) = \max_{i,j} D(i,j)$$

在公式中，当$\lambda_i = 1, p_1 = 1, p_{j \neq 1} = 0$时，此集成方案就是所谓的投票法（Voting）。

多分类器组合方法：

投票法 1(等权法)：每个分量识别器的权值都相等；

投票法 2(最好占优法)：做实验选到 5 分量识别器的恰当的权重；

表 8-7 字母识别率（%）

	基于特征 1	**基于特征 2**	**基于特征 3**	**基于特征 4**	**基于特征 5**	**综合识别率**
单独识别率	51.57	50.74	51.73	43.05	62.29	
权重（投票法 1）	0.2	0.2	0.2	0.2	0.2	70.19
权重（投票法 2）	0.14	0.2	0.14	0.18	0.34	75.06

8.4.1.4 基于卷积神经网络的全形态字母识别

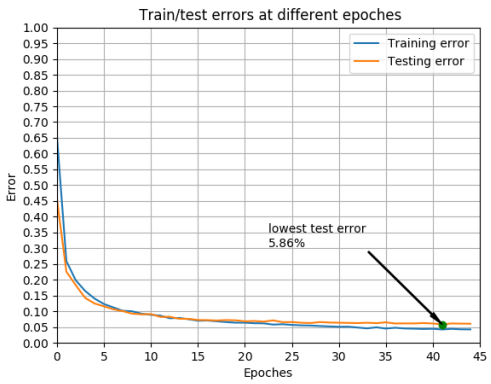
卷积核即基本形状单元滤波器核的大小为 3*3。网络中底层卷积提取低级特征或局部特征，高层中的卷积则对应高层特征或全局特征。在低层卷积使用较少的滤波器，而在高层中的卷积中设置了更多的滤波器。在卷积操作之后通过补零的技术保持了特征图大小不变，这有助于计算和增加网络层数。池化操作用取最大值池化把特征图大小减半，采用了 2*2 的池化区域和 2 步伐。全连接层对自动学习的特征进行全局组合以后通过 softmax 得出最后的分类结果。它由 5 个卷积层和 5 个池化层和 3 个全连接层组成。

如果设定的归一化样本尺寸小，则原始样本归一化之后丢失很多信息。归一化尺寸较大会保留很多原始样本信息，但需要很大的存储空间。本文,把联机手写字母样本首先归一化到 46*46 的尺寸，然后用外围补零的方法增大到 48*48。用此尺寸，可以采用 5 个卷积层。表 8-8 记录了用此尺寸和具有 4 个和 5 个卷积层的卷积神经网络实验结果。可以看出，较大的归一化图像和较深的网络结构给出了更好的识别结果。

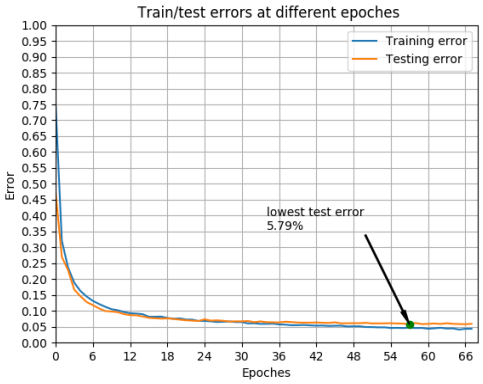
表 8-8 基于 48*48 字母图像的卷积神经网络训练记录

卷积层数	网络结构	模型大小	停止 epoch	训练集上识别错误率	测试集上最小识别错误率	测试集上识别错误率
4Conv	[C24-P2-d0.2]-[C48-P2-d0.2]-[C96-p2-D0.3]-[C192-P2-D0.3]-[FC256-D0.4]-[FC256-D0.4-FC128]-Softmax128	8.99 M	45	4.28%	5.86%	6.14%
5Conv	[C24-P2-d0.2]-[C48-P2-d0.2]-[C96-p2-D0.3]-[C192-P2-D0.3]-[C384-p2-D0.3]-----[FC256-D0.4]-[FC256-D0.4-FC128]-Softmax128	12.26 M	68	4.36%	5.79%	5.83%

如图 8-24 所示，表 8-8 中的两个卷积神经网络模型在识别率和模型泛化性能上都得到了改善。除了正则化和模型结构的贡献外，较大图像保留的信息有助于识别性能的提高。卷积神经网络的初步训练实验中得到的结果也很好。包含 5 个卷积层的卷积神经网络在 48*48 输入图像上有效识别错误率降到 5.83%。此识别率是在 10240 个测试样本上进行三次识别测试的平均错误率，相当于 94.17%的识别准确率。



(1) 4 层卷积 CNN 模型



(2) 5 层卷积 CNN 模型

图 8-24 48*48CNN 字母识别模型训练实验

8.4.2 联机手写维吾尔文单词分割

切分技术的划分结构大致可以按图 8-25 所示。常用的字符串切分方法主要分为两种，显式切分（explicit segmentation）和隐式切分（implicit segmentation）两种。根据在字符切分中的单词分割和字符识别之间的关系，目前显示切分又可以分为基于识别的切分和无切分识别。

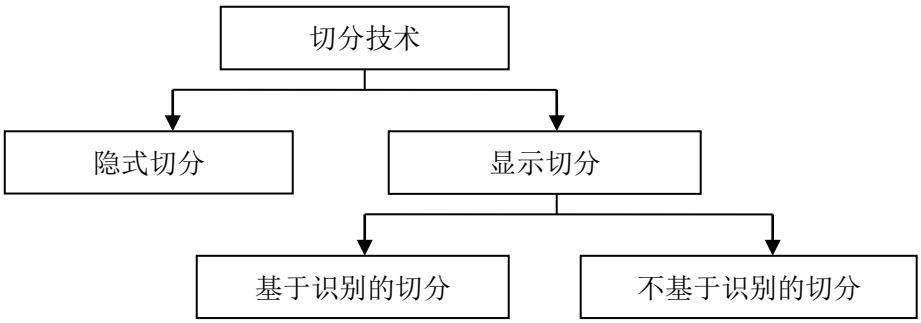


图 8-25 切分方法分类图

8.4.2.1 过切分算法

该算法的流程图如图 8-26 所示。

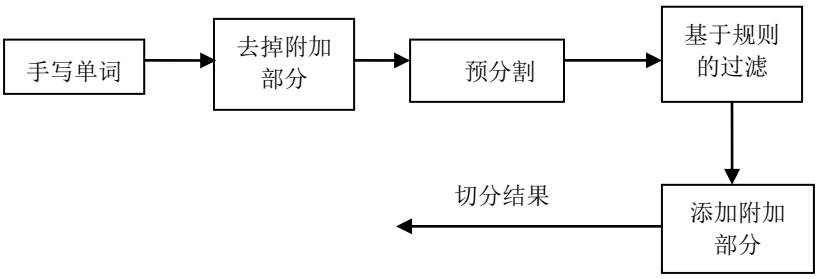


图 8-26 过切分算法流程图

第一步：去掉附加部分

在手写维吾尔单词过程中，笔尖运动轨迹由它在手写板上的x, y坐标和“落笔”、“抬笔”的状态来描述，将笔尖在手写板上的运动轨迹分隔为笔划序列。不管笔划的书写顺序，每个单词包括的笔划可分为主笔划和附加笔划。所以，笔划可能代表连体段的主要部分，也可能是个单个字母的主要部分，或有时甚至是个一个点。检测并去掉附加部分是个重要步骤。为了区分维吾尔单词的主要部分和附加部分，同时满足以下条件：第一，对每个笔划进行计算下面的几何特征： 每个笔划的宽度、高度、宽高比。如果这些值小于预定的阈值，那么这个笔划属于附加部分。第二，如果这个笔画的书写方向为从右到左（$x_i - x_{i+1}$>0），那么这个笔划属于附加部分。如图 8-27 (a)所示原始单词图像，如图 8-27(b)所示去掉附加部分后的图像。

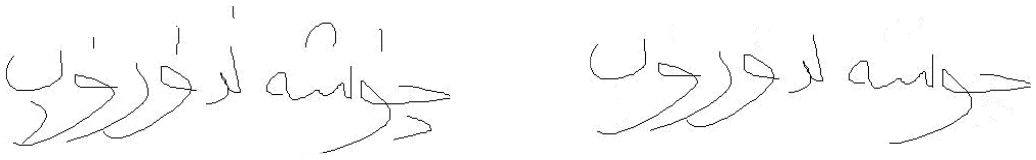


图 8-27 (a)原始单词图像; (b)去掉附加部分后的图像

第二步：预切分

生成候选分割路径的基本步骤为：

确定候选分割点。计算从每一个坐标序列点P_i到它的下一个点

P_{i+1}的倾斜角度（$\overline{P_iP_{i+1}}$与水平线之间的角度）。如果这个角度α小于$\frac{\pi}{6}$（实验参数值）并且书写方向为从右到左（$x_i - x_{i+1}$>0），则该坐标点被称为候选分割点。

处理重叠。利用空间信息将删除一些不满足条件的错误候选分割点。如图 8-28 所示，如果过某个初分割点的直线与从(90-α)o 到(90+α)o 的孤上任有一个交叉点(也可以说，如果从(90-α)o 到(90+α)o 的范围内有覆盖点)，则从初步分割点组中删除这个分割点并其余的点都保留。α是经验值 25o。

图 8-28 (a)该分割点被删除; (b)该分割点被接受

生成分割线段。通过连接成连续的分割点形成切分线段并确定最终分割点。计算每两个初始分离点之间的水平距离，如果每两个相邻初步分割点差$g_i - g_{i+1}$<8 (实验经验值)，则连接成分割点形成切分线段，否则该两个点属于不同的切分线段，如图 8-29（a）所示。

定位分割点。以上步骤中，本算法可以找到 K 个切分线段。每个切分线段的中间位置就是被判断为分割点。 从而得到最终的分割点系列S_i (i=1, 2,⋯, k), 如图 8-29（b）所示。

图 8-29 (a)切分线段; (b)最终过分割结果

第三步：基于规则的过滤

在本步骤中，通过基于规则的方法过滤删除一些不满足条件的

错误分割点。规则 1：根据基线删除多余的分割点。先根据主笔画点序列计算该主笔画的基线，然后上一步确定的切分点中删除不在基线和基线附近的分割点。这里所提到的基线是指对主笔画点序列进行水平投影后，投影值最大的那条线。如果每个分割点对应的纵坐标 Y 值与基线值的差大于 10（实验参数值），则判断为该分割点不是正确的分割点并删除。规则 2：如果两个被建议的分割点之间的距离小于一个预定义的阈值（5 经验），那么删除该分割点。

第四步：添加附加部分

分割点被检测以后，需要重新把附加部分分配给所属于的切分块，一般这些切分块是字母的主体部分，或是主体部分相连的段。附加部分的正确合并对于切分块识别率的影响非常大。如果不把附加部分分配给这些主体笔划，那么有些原切分块不能单独构成一个字母，因为很多字母的主体笔划是相似或是相同的，唯一区分它们的是它们的附加部分。本文通过计算附加部分和切分块之间的重叠度来判定附加部分的归属。重叠度是利用它们边界框的大小和位置信息计算得到。假设边界框用上下左右边界的坐标表示，那么附加部分的边界框表示为$(x_d^l, x_d^r, x_d^t, x_d^b)$，切分块的边界框表示为$(x_s^l, x_s^r, x_s^t, x_s^b)$。假定$x_s^l$<$x_d^l$，如果$x_d^l$<$x_s^r$，那么它们相重叠。重叠度和跨度分别表示如下：

$$overlap = x_s^r - x_d^l$$

$$\text{span}=\max(x_s^r, x_d^r) - x_s^l$$

归一化的重叠度计算如下：

$$normOverlap = \frac{1}{2}\left(\frac{overlap}{width_1} + \frac{overlap}{width_2}\right) - \frac{dist}{span}$$

其中 width1 和 width2 分别表示的宽度，dist 表示它们中心的水平距离。

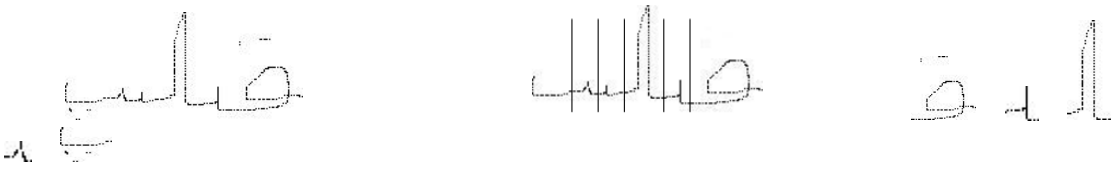


图 8-30 过切分结果

(a)由 5 个附加笔画组成的一个单词 (b) 去掉附加部分后切分结果

(c) 重建附加部分的结果

过分割方法在不同手写维吾尔单词数据集上的测试性能见表 8-9。从表中可见，过分割点总数比正确分割点总数超过了 40%-50%；召回率，精度，F 度量分别达到 98%，56% 和 71%。图 8-30 中给出了过分割后的部分结果图。

表 8-9 过分割方法在不同数据集上性能

	正确分割点总数	过分割点总数	召回率 (%)	精度 (%)	F 度量
数据集 1	1875	3298	98.78	56.85	71.54
数据集 2	1832	3465	98.26	52.87	68.81
数据集 3	1865	3750	96.52	49.73	65.75

8.4.2.2 基于动态规划的联机手写单词分割方法

利用动态规划方法解决切分问题的关键在于代价函数的设计，该函数用于描述将一个或多个切分块进行合并时所需的代价。由于本文针对的是维吾尔文单词的切分，因此由每条路径得出的候选切分块，送入维吾尔文字母识别器进行识别后，识别器将给出识别结果。根据切分块的识别信息，得到最终对应整个单词的代价函数。

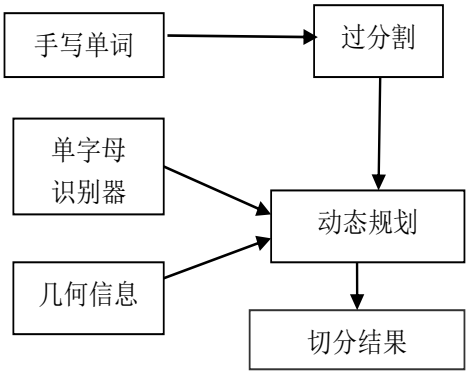


图 8-31 联机手写单词分割算法流程图

经过上述的过分割过程中，可以得到一系列切分点以及对应的字母基本片段序列，分别用$\{P_0,\mathrm{P}_1,\ldots,\mathrm{P}_N\}$和$\{S_0,\mathrm{S}_1,\ldots,\mathrm{S}_{\mathrm{N}+1}\}$来表示。过分割的结果往往具有一定的冗余度，过分割后字符可能会被切分成多个部分。一些基本片段可能包含一个单一的维吾尔文字母，还有一些可能只包含字母的一部分，需要做进一步的合并。根据过分割结果构造一个切分候选网格。有候选字符模式构成切分候选网格，网格中从起点到终点的一条路径就对应单词的一种切分方式。

表 8-10 中给出了结合考虑单字母识别信息和几何信息的分割方法在不同数据集上的实验结果。在代价函数中引入基于切分块的几何信息，并与单字母识别器信息进行线性加权，以提高分割性能。实验结果表明，该方法对不同大小的手写字母都能得到满意的效果。

表 8-10 结合单字母识别信息和几何信息的分割方法在不同数据集上性能

	召回率 (%)	**精度 (%)**	**F 度量**
数据集 1	95.66	73.54	82.76
数据集 2	94.38	70.18	80.77
数据集 3	93.49	67.37	78.24

8.4.3 联机手写维吾尔文单词识别

根据切分、识别和语义信息利用方法，可以把字符串识别方法分为如图 8-32 所示的几类。无切分识别方法不需要识别出每个字符，而是直接对整个单词进行识别，从而避免了切割。该方法从语音识别发展过来，要对每一个单词类别进行建模，因此这种方法只适用于小词典的情况。基于切分的识别方法，则是将字符串识别看成是字符识别的串联，因此可以适用于字典驱动的和无字典驱动的字符串识别，并能应用到大类别集的字符串识别问题。

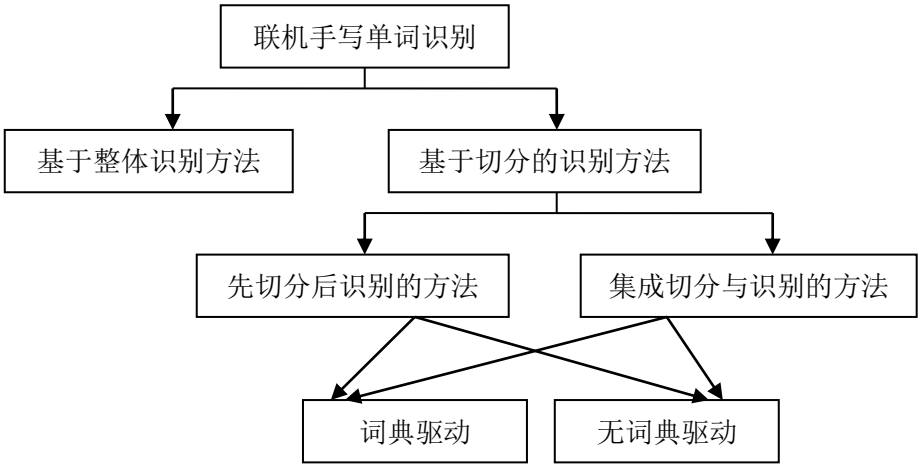


图 8-32 联机手写单词识别的方法分类图

8.4.3.1 基于词典驱动的单词识别方法

基于词典驱动的联机手写维吾尔文单词识别系统结构如图 8-33 所示，它包含输入单词图像和词汇词典两个输入（输入图像为图像文件，词汇词典为文本文档），系统输出就是识别结果。该系统由两大部分组成，一个是切分部分，另一个是匹配部分。在前一个部分中，连续的片段动态的合并为候选字符模式，不同的合并方式产生不同的候选字符序列，这样可以构成一个切分候选网格。在后一个部分中，结合维吾尔语词典匹配来寻找到一条最优路径作为识别结果。整个系统主要包含 5 个大的功能模块：预处理模块，单词图像过切分模块，

构造切分候选网格模块，单字母识别器模块，路径匹配模块。

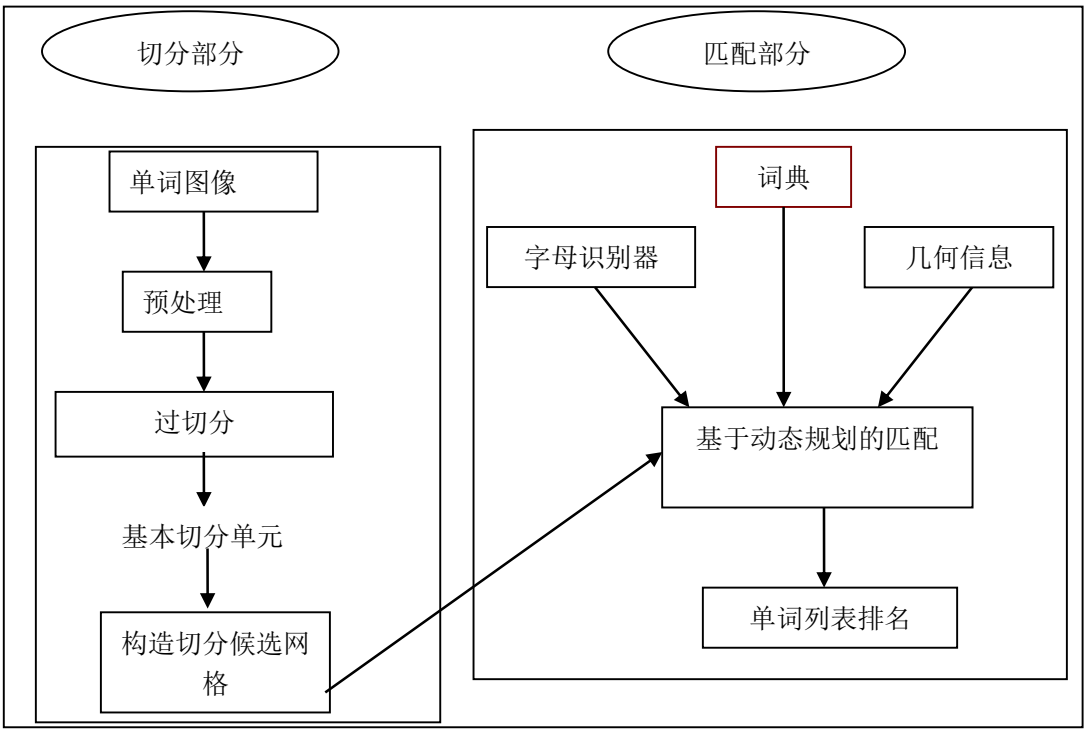


图 8-33 维吾尔文单词识别系统框架

问题描述

在过切分模块中，根据一定的规则将对预处理后的单词图像进行初步的切分，得到的只是预切分出的基本单元序列。输入单词图像 T 被过切分之后，可以表示一个从左至右排列的有序过切分片段序列：

$$T = \{s_0, s_1, \ldots, s_{S_N-1}\}$$

其中S_N 表示输入单词图像被分割成的所有过切分片段的总数，S_k 表示第 k 个过切分片段。而且一个过切分片段可能包含一个单一的维吾尔文字母，或者可能只包含字母的一部分。

由维吾尔文单词组成的词典R_i可以表示为：

$$R_i = \{c_i(0), c_i(1), \ldots, c_i(C_N(i)-1)\} \quad 1 \le i \le L_N$$

其中$C_N(i)$表示在词典中第 i 个词包含的字母数，L_N是词典条目的数量，和$c_i(j)$表示词典中第 i 个词的第 j 个字母。

单词包含的所有字母的内码C_m：

$$C_m = \{X(i)\}_{i=1}^{m}$$

其中 m=1，….,128 代表128个字母，$X(i)$表示第i个字母的内码。

相邻的基本单元可能进一步合并，最终成为切分获得的维吾尔文字母。过切分基本片段的合并过程定义如下：

$$S(b,e) = \{s_b \oplus s_{b+1} \oplus \ldots \oplus s_e\} = \{s_v\}_{v=b}^{e}$$
$$0 \le b \le S_N, 0 \le e \le S_N, b \le e$$

其中 b 是切分开始位置和 e 是切分结束位置，$\oplus$是合并操作符。

一个输入单词图像 T 与对应的词典单词之间的一个映射被定义为：

$$\Psi = \{(c_i(0), s_0 \oplus s_r), (c_i(1), s_{r+1} \oplus s_k), \ldots, (c_i(C_N(i)-1), s_q \oplus s_{S_N-1})\}$$

其中$s_{r+1} \oplus s_k$是一个过切分片段子序列，在匹配过程中该子序列被组合为一个候选字母模式。所以，单词识别问题可以被描述为输入单词图像的过切分片段序列和词典中的词条之间的最优匹配。也就是，为输入单词图像 T，将过切分片段动态的分配给每一个词典中的词条，可以表示为如下：

$$D^* = \min_{1 \le i \le L_N} D(R_i, T)$$

路径匹配

在第一阶段中，在一个给定的词典条目包含的每个字母$(c_i(j), 0 < j < c_N(i))$与输入单词图像 T 的任意切分子序列部分（b 开始， e 结束）之间计算匹配距离。

$$\hat{D}(v,b,e) = d\left(F(S(b,e)), X(v)\right) \quad \text{for } 0 \le v \le c_i(j)$$

其中$F(.)$表示对应切分块的特征提取函数。$\mathrm{d}(\cdot,\ \cdot)$是两个特征向量之间的匹配距离（代价函数），它来自于字母分类器和几何模型输出的置信度值的和。

为了使用动态规划算法寻找最优匹配，我们定义$\overline{D}(b,e)$作为词典中每一个词条R_i和输入单词图像过切分片段序列$T = \{s_0, s_1, \ldots, s_{S_N-1}\}$的最优匹配累加代价。从以下公式中可以得到对某种切分候选路径得

到的最佳匹配距离，本文中采用归一化编辑距离（Normalized Edit Distance）。

$$\overline{D}(b,e)=\min_{1\le v\le c_i(j)}\left[\frac{1}{c_i(j)}\hat{D}(v,b,e)\right]$$

定义输入单词图像$T=\{s_0,s_1,\ldots,s_{S_N-1}\}$最多可能切分成$s_N$个切分片段，词典每一个词条$R_i=\{c_i(0),c_i(1),\ldots,c_i(C_N(i)-1)\}$包含的字母数为$C_N(i)$。

而且，词典中某个词包含的字母和过切分图像片段之间的关系应该满足条件$C_N(i)<S_N$。也就是说，如果一个词典中的词条不满足该条件，那么该词条R_i被拒绝，从而得到限制匹配范围的目的，它不会影响识别的准确性。

在第二阶段中，通过第一阶段得到的单个匹配分数相结合，计算整个词典条目的累计匹配代价。在最佳路径中，每第j个字母对应的，e切分点结尾切分块的动态规划距离的计算方法是如下：

$$\overline{D}_j(e)=\min_{1\le b\le e}\left[\overline{D}(b,e)+\overline{D}_j(b-1)\right]$$

为了决定最佳路径，根据以上公式中的递归过程，我们可以制定动态匹配过程的第二个阶段的算法描述为：

步骤1： 初始化

$$\overline{D}_0(e)=\overline{D}(0,e)\qquad 0\le e\le M_d-1$$

步骤2：递归

$$\overline{D}_1(e)=\min_b\left[\overline{D}(b,e)+\overline{D}_0(b-1)\right]\qquad e_{min}^{1_{max}^{1}}$$

$$\overline{D}_2(e)=\min_b\left[\overline{D}(b,e)+\overline{D}_1(b-1)\right]\qquad e_{min}^{2_{max}^{2}}$$

•••

$$\overline{D}_j(e)=\min_b\left[\overline{D}(b,e)+\overline{D}_{j-1}(b-1)\right]\qquad e_{min}^{j_{max}^{j}}$$

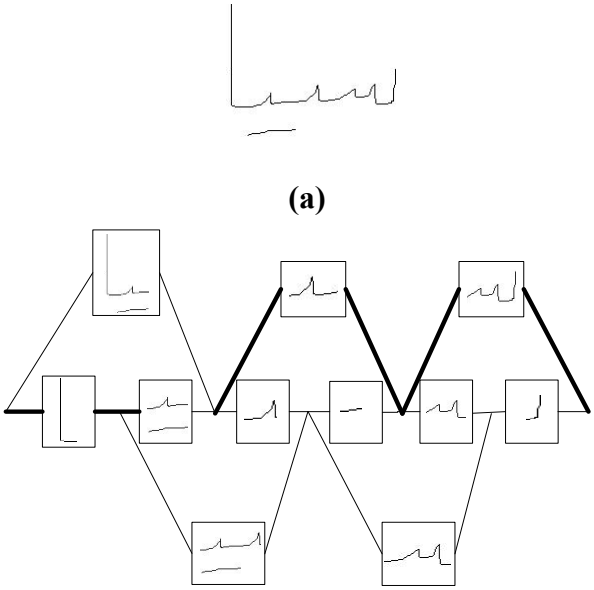
•••

$$\overline{D}_{C_N-1}(e)=\min_b\left[\overline{D}(b,e)+\overline{D}_{C_N-2}(b-1)\right]\qquad e_{min}^{C_N-1_{max}^{C_N-1}}$$

步骤3： 最终最小距离

$$D^*-\overline{D}_{C_N-1}(S_N-1)$$

首先采用过切分算法产生一系列切分基本片段，然后根据过切分结果，通过合并相邻的切分基元得到的切分候选网格，如图 8-34 所示。通过驱动词典信息，由相邻切分基元组成的候选字母模式，通过字符分类器进行分类识别，得到其置信度得分。把这些候选字符模式的置信度得分联合起来，得到整个单词的切分方式和类别序列的置信度得分。然后通过动态规划方法，将识别信息与词典信息集成进来，搜索最优的路径，也就是从所有可能的情况中选择最优的路径作为最终的识别结果。



(a)

(b)

图 8-34 根据过切分结果构造切分候选网格的示意图

（粗黑线表示最优的路径）

每一种匹配方式对应一条从左到右的切分路径，为了寻找最优的匹配，每条路径将用一个代价函数来表示。加权参数分别设置为 λ_1 =0.9, λ_2=0.1。在表 8-11 中，给出联机维吾尔文手写单词识别率。代价函数是动态规划算法中比较重要的一个因素，它的选取将影响到最小代价路径的搜索结果。在确定了代价函数之后，就可以利用动态规划的方法从单词的开始分割块到结束分割块搜索出具有最小代价的

分割路径。用到两种距离度量方法，即求和编辑距离和归一化编辑距离。实验中用的词典大小分别为 100,500,1000,10000。对应不同大小的词典，维吾尔文单词识别率分别为 84%、78%、68% 和 46%。

表 8-11 联机维吾尔文手写单词识别率

词典大小		100	500	1000	10,000
求和距离	首选	74%	64%	54%	32%
	前 10	86%	80%	68%	48%
归一化距离	首选	84%	78%	68%	46%
	前 10	96%	90%	78%	58%

8.4.3.2 基于循环神经网络的端到端联机手写单词识别

采用的RNN和CTC的结合模型很好的解决维吾尔文单词识别中的字母分割问题，直接将手写单词轨迹映射到维吾尔文字符串中，也不需要做专门的字母识别器来识别分割出来的字母候选轨迹段。在图8-35中显示了基于RNN和CTC的端到端词典内容无约束的单词识别系统。

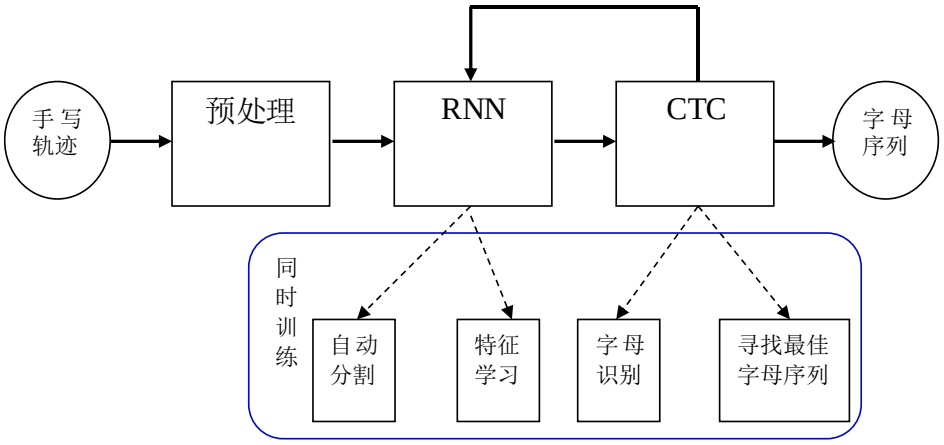


图 8-35 基于 RNN+CTC 的端到端无约束联机手写单词识别框架

通过 RNN 和 CTC 的结合模型可以获取字母级别的输出序列，实现了词典内容无约束的手写单词识别系统。本文用 RNN 和 CTC 的结合模型作为本文无约束手写单词识别基础模型，如图 8-36 所示。基础模型使用两个双向循环网络层，三个一维取平均池化层，两个全连接层，最后全连接层输出直接传给链接主义时序分类 CTC 解码器。一维池化操作有利于提高特征鲁棒性和对特征序列的缩短。CTC 解码器生成字母序列作为直接产生字母级别识别结果，及字母标签序列。

模型最后全连接层每个时刻的输出通过 Softmax 激活函数分别投影到特定的标签概率上。CTC 使用一个特殊设置的空白标签，表示没有输入或还没有收到能形成字母的序列信息。物理意义上近似可以理解为手写轨迹中的断点。因此，最后一个 FC 层设置了 n+1 个节点。模型最后一个全连接层的神经元数量由字母表中的字母/字母数决定，即默认为 128 和一个专门为 CTC 解码而设计的空白标签来设置。所以，本文基础 RNN 模型中最后输出层上的节点数维 128+1。CTC 具有匹配可变长度输入-输出序列的能力。它可以学习可变长度的时间序列特征和直接生成字母标签序列。模型输出是手写轨迹可能包含的字母串（字母序列），由 CTC 解码器通过束搜索算法得到。本文使用束搜索算法从候选路径中选择最佳路径和响应的标签序列作为输出。采用默认的束搜宽度为 100。在实验中仅用最佳标签序列来进行评估。

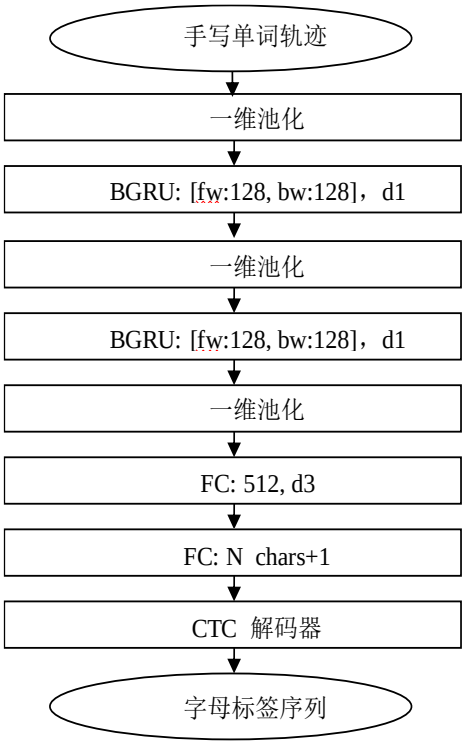


图 8-36 RNN-CTC 无约束联机手写维吾尔文单词识别模型结构

（fw，bw 分别表示前向和后向）

表 8-12 平均模型和拼接模型识别性能比较 CAR（%）

模型	#Ep	T/ep（min）	R-S（s）	Train-CAR	Test-CAR	OUT1-CAR	OUT2-CAR
Char128 拼接	56	~51	0.039	92.52	89.99	45.63	29.59
Char128 平均	44	~45	0.035	96.94	93.39	69.45	46.96

表中，T/ep 表示每个训练期需要的平均时间，R-S 表示识别单个单词需要的平均时间，Train- CAR 和 Test-CAR 分别是训练集上的编辑字符准确率，OUT1-CAR 和 OUT2-CAR 分别是集外词测试集 HW.OUT1 和 HW.OUT2 上的编辑字符准确率。

表 8-13 Char34 和 Char128 模型识别性能比较 CAR（%）

模型	#Ep	T/ep	R-S	Train-	Test-	OUT1-	OUT2-

		（min）	（s）	CAR	CAR	CAR	CAR
Char128 平均	44	~45	0.035	96.94	93.39	69.45	46.96
Char34 平均	40	~35	0.031	96.64	93.32	**73.24**	**52.06**

表中，T/ep 表示每个训练期需要的平均时间，R-S 表示识别单个单词需要的平均时间，Train- CAR 和 Test-CAR 分别是训练集上的编辑字符准确率，OUT1-CAR 和 OUT2-CAR 分别是集外词测试集 HW.OUT1 和 HW.OUT2 上的编辑字符准确率。

8.4.3.3 基于一维卷积和循环神经网络的无约束单词识别

在结合模型中，一维卷积层用于基本序列特征提取，循环网络层用于上下文信息提取。在每两个连续的一维卷积层之后，用批归一化来改进网络层之间参数值的识别性能和下溢。由于循环运算比卷积运算要慢得多，因此该模型只保留了一个循环网络层。

在实验中对比的不同卷积层组结构在表 8-14 给出。其中，符号 C 表示一维卷积层。例如，C[32,32]表示两个连续的一维卷积层，每层有 32 个单元。一维卷积使用内核大小为 3 和步长为 1 进行操作；P 表示池化内核为 2 和步长为 2 的一维平均池化操作。本章对比实验中的基础模型和结合模型的网络参数都约在 2.5M 左右。在每个卷积层组中放了相同宽度的两个连续卷积层，即第一个卷积层的数据直接输入到第二个卷积层进行特征分析。第二个卷积层的数据作为该卷积层组的输出进行下一步处理，比如通过一维池化过程或直接送到下一个卷积层组等。最后一个卷积层组的输出通过一维池化处理后输入到循环网络层。根据一维卷积（1D-Conv）层组的网络结构，本章中进行比较的一维卷积和循环网络结合不同模型结构起了简单的名字，比如 1D-Conv-3-E、1D-Conv-3 和 1D-Conv-4 模型。

表 8-14 结合模型一维卷积组结构

数据	一维卷积组结构

1D-Conv -3-E	C[128,128]P-C[128,128]P-C[128,128]P
1D-Conv -3	C[64,64]P -C[128,128]P-C[256,256]P
1D-Conv -4:	C[32,32]- C[64,64]P-C[128,128]P -C[256,256]P

1D-Conv-3-E 和 1D-Conv-3 模型有三个卷积层组。1D-Conv-3-E（ E 表示 Equal,同等的意思)中的三个卷积层组都用了同等的宽度，即每个卷积层都设置了 128 个节点。1D-Conv-3 模型中三个卷积层组宽度不同。三个卷积层组的宽度分别为 64、128 和 256，用了从窄到款的结构。因为底层的卷积被认为学习输入序列中的基本特征，高层的卷积队基本特用更多维度综合。1D-Conv-4 模型分别有四个卷积层组，其中底层卷积组中的卷积层宽度比高层卷积组的宽度小。模型结构设计的时候，我尽量保持了模型总参数量保持一致。

一维卷积和循环网络结合模型的识别性能在表 8-15 中进行对比。可以发现一维卷积模型对识别率的提高有明显的贡献。同时，还可以发现更多有用的信息。

表 8-15 一维卷积结合模型识别实验结果 CAR（%）

模型	#Ep	T/ep（min）	RecS（s）	Train CAR	Test CAR	OUT1 CAR	OUT2 CAR
基础模型	40	~35	0.031	96.64	93.32	**73.24**	**52.06**
1D-Conv -3-E	34	~24	0.018	96.92	94.95	78.29	56.47
1D-Conv -3	34	~32	0.031	94.87	91.53	82.65	63.27
1D-Conv -4	36	~31	0.027	96.98	93.21	83.23	63.56

表中，T/ep 表示每个训练期需要的平均时间，RecS 表示识别单个单词需要的平均时间，Train- CAR 和 Test-CAR 分别是训练集上的编辑字符准确率，OUT1-CAR 和 OUT2-CAR 分别是集外词测试集 HW.OUT1 和 HW.OUT2 上的编辑字符准确率。

8.4.3.4 基于数据增强的无约束联机手单词识别

数据量的大小直接联系到数据的表示能力，收集的数据量越大能包含的样本变化越多，越接近于实际情况。在手写识别研究中，收集大量手写样本往往需要的大量的人力和财力，是一个很困难而且漫长的过程。手写数据增强（Data Augmentation）用少量的原始手写数据来构造更多的伪造样本，从而增加数据量并提高数据表示能力，是减轻或弥补数据缺少问题的一种有效途径。

一种方法直接应用于整体手写单词轨迹会导致不理想的结果。根据维吾尔文单词的手写特性，借鉴脱机和联机手写数据增强方法，本文提出多种手写增强算法结合的方法，如图 8-37 中所示。按照不同数据增强方法的优缺点，本文提出或采用的数据增强算法分别在个别笔画和整体样本上实现。

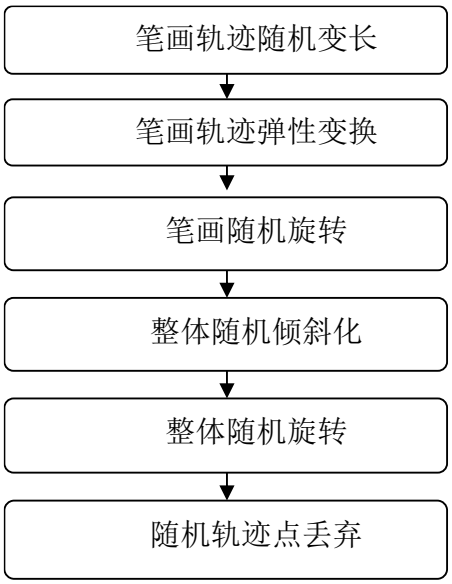
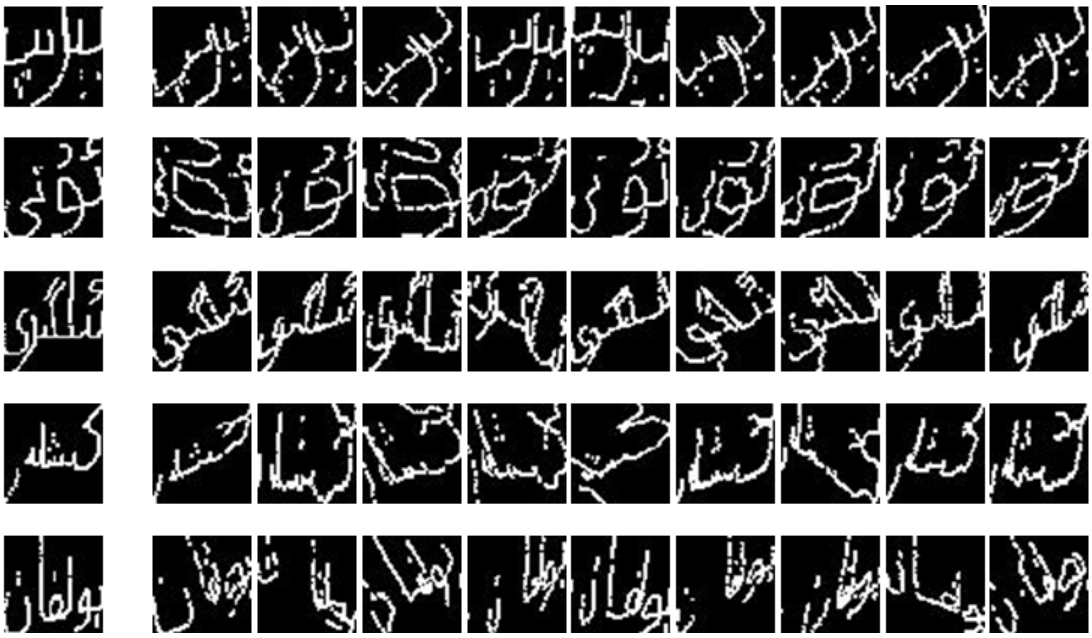


图 8-37 多种数据增强算法结合流程

从图中可以看出，这种结果提供了用非常少的原始样本来构造具有不同手写风格的更多伪造样本，大幅度提高了手写数据增强效果。更多手写维吾尔文单词上的数据增强效果在图 8-38 中显示。



（a）原始样本　　（b）数据增强以后的伪造样本

图 8-38 用数据增强构造的手写单词样式

用伪造数据放在训练后，在集外词测试集上的识别结果也有所改进。常用词的集外词 OUT1 上的识别结果比仅用原始数据时的 CAR 结果提高了 0.79 个百分点。当训练集数据量增加到原始数据量 3 倍和 4 倍时，识别率获得了进一步提高，份别比用原始数据的编辑字母准确率提高了 1.28 和 1.89 个百分点。在 OUT1 获得的最高编辑字母准确率达到了 85.12%。每增加一杯伪造样本时，OUT1 上的识别率幅度也开始有错下降。主要原因随着伪造样本在训练集中的占比的增大，深度神经网络模型开始倾向于伪造样本的风格。需要注意的是，测试集中的手写单词样本都是原始的样本，没有伪造样本放在测试集里。

表 8-16 基于数据增强的单词识别结果 CAR（%）

模型	训练数据量	Train-CAR	Test-CAR	OUT1-CAR	OUT2-CAR
原始数据	99120	96.98	93.21	**83.23**	**63.56**
+构造数据	+99120	96.62	95.36	84.02	63.59
+构造数据	+2*99120	96.56	95.86	84.51	63.42
+构造数据	+3*99120	96.46	95.98	85.12	63.01

在外来词测试集 OUT2 上发现与以上集内词和常用词集外词识别的情况不太一样的现象。在 OUT2 上的识别结果没有观察到明显的变化。当训练集数据量增加到原始数据量 3 倍和 4 倍时，外来词识别率没有提高，反而出现了小幅度下降。这主要是因为，构造出来的数据都是基于集内词样本进构造的。这导致模型更倾向于集内词的识别。

本章编写人员：

艾斯卡尔·艾木都拉、米吉提·阿不里米提、玛依热·依布拉音

第 9 章 国际中文教育智能处理

随着时代的进步和学科的发展，技术的作用已遍及并深入汉语教学实践和教学研究的方方面面，技术应用已成为汉语教学的基本特征。从汉语各要素、各技能教学，到课堂教学模式和教学活动、汉语测试，以及教师培训和教师培养等，都可以看到技术的身影。从录音/录像技术、到广播/电视技术、再到多媒体技术和网络技术，汉语国际教育领域中技术应用的路径始终合着新兴技术在教育领域应用的脉搏，取得了丰硕的成果[12]。当下智能技术的发展更是极大促进了汉语国际教育的革新。本章将从教学资源、教学实践、中文测试、语料库、综合应用平台五个方面探讨智能技术在国际中文教育中的应用和影响。

9.1 智能技术赋能教学资源开发

国际中文教育资源是中文和中国文化“走出去”的重要载体，中文教学资源建设是国际中文教育事业发展的重要内容。近年来，智能技术逐渐渗入国际中文教育领域，语音识别、文字识别、自然语言处理、虚拟现实等技术在国际中文教育领域赋能越来越丰富的应用。中文数字资源的“技术含量”越来越高，智能化产品越来越丰富，为加速并深化信息技术与国际中文教育的融合发展奠定了基础[366]。本节将对智能技术赋能下的国际中文教育资源开发进行探讨。

9.1.1 多媒体技术

传统媒体时代，中文教育资源因其语言文字的属性多以纯文字的文本形式存在。这一时期的教材以文本、图形、图像等无交互特性的静态媒体作为主要的内容形式。随着多模态教学理论的提出与发展，多模态资源正被越来越多地用于国际中文教育。对语言学习来说，多

[12] 郑艳群.汉语教学 70 年——教育技术的影响及作用[J].国际汉语教学研究,2019(04):69-76.

模态资源加强了感官刺激，丰富了情感体验，提高了语言学习的趣味性；多模态资源的不同模态之间具有互文性，有助于学习者更加准确地把握语境信息、理解文化背景，提高对语言形式的敏感度；借助多模态资源不同模态的互补性，在不减少文本内容信息的前提下降低资源中文本的比重和难度，可以降低学习者（尤其是初学者）参与中文真实交际的门槛，增加其在交际中提升中文运用能力的机会。

语言教学往往依赖于多媒体呈现方式，传统纸质媒介教学资源有限的承载能力无法满足语言教学的多模态需求，智能化多媒体技术对国际中文教材和课堂教学都产生了很大的影响。

1. 多媒体技术对教材的影响

多媒体技术促成了数字教材的诞生和发展。数字教材也称为电子教材，是依托于信息技术开发、超越时空的多媒体教材，具有字、音、形、色、义等的合成性、动态性及可再生性等特点[367]。能够提供音频、视频、图像、动画等资源，类型较为丰富，也可以为学习者提供交互，增加沉浸感。代表教材主要有《新实用汉语课本》（北京语言大学出版社）、《发展汉语》（北京语言大学出版社）、《长城汉语》（第2版）（外语教学与研究出版社）、《快乐汉语》（人民教育出版社）等。

数字教材的雏形是纸质教材出版同时附带电子音像产品，最终目标是“全媒体出版”。“全媒体出版”是指出版物既以传统方式进行纸媒印刷品发行又以数字出版物的形式通过互联网平台、无线阅读平台以及阅读器等数字设备进行同步发行[368]。网络的逐渐普及、平板电脑及智能手机的广泛使用助推了全媒体出版的步伐，同时也进一步催生了大量全媒体汉语学习资源的出现，如《写汉字》（Chinese Writer）、《汉语拼音发声器》（Chinese Pinyin Audio Trainer）、《中文学习工具》（Chinese Dictionary Flashcards）等。

2. 多媒体技术对教学的影响

多媒体辅助课堂教学是较为传统的辅助教学手段，其研究主要集

中在多媒体技术在教学实践中的运用。如王飙[369]研究了影视作品和美术片在不同阶段的视听说教学中的差异；吴双[370]研究了不同多媒体技术在写作前期导入话题、中期写作练习、后期作文讲评中的应用；汝淑媛[371]介绍了如何将多媒体技术应用于语法意义的展示、语法规则的练习、语法知识的综合运用三个教学阶段；陈新[372]基于多模态理论框架设计了汉语视听说教学模式，对软硬件设施配置、教学流程等提出了具体的要求。多媒体辅助课堂教学的应用成果主要有教学手册、教学资源库等。

智能技术的引入对多媒体辅助课堂教学也产生了影响，尤其在教学效果评估方面。多种新技术正在被用于教学效果的量化评估，如刘汉钦[373]借助眼动技术探讨了教学多媒体对汉语二语学习者汉字认知的影响。

除了课堂教学外，影音技术和互联网技术的结合还促成了慕课、微课等全新教学模式的发展。

慕课是基于互联网产生的大规模开放式课程，慕课课程整合多种社交网络工具和多种形式的数字化资源，形成多元化的学习工具和丰富的课程资源。在国际中文教育领域，慕课的价值与影响伴随中国语言与文化国际传播数字化进程的加快而不断彰显。从课程类型来看，慕课可分为四大类：语言学习类、文化学习类、教师发展类和专业学习类，其中语言学习类课程最为丰富。从学习内容来看，中文教学慕课可分为综合类、语言要素类（语音、词汇、语法、汉字）、语言技能类（听、说、读、写）类、考试类和专门用途中文类，其中综合类慕课数量最多，占半数以上[374]。

微课是指依据认知规律、运用信息技术，支持碎片化学习而呈现的学习内容及扩展素材的结构化数字资源。微课主要集中发布在公开微课赛事、教学平台、视频网站上。相关赛事主要有“全国研究生汉语教学微课大赛”、“全国高校汉语国际教育专业研究生教学技能大赛”

等；教学平台主要有全球中文学习平台、“中文联盟”数字化云平台、唐风汉语国际教育云平台等；国内外视频网站主要有 bilibili、YouTube、抖音等。

9.1.2 数字化交互技术

数字化教学初期的教学资源多为“展示型”，通常只是把纸质版本的资源转化为电子资源，并没有改变学生被动接受的学习模式。但随着教学理念的革新和智能技术的发展，“交互型”学习资源成为新的发展趋势。

学习者与学习内容的交互通常被理解为学习者浏览阅读各种类型的学习材料的过程。当学习材料中的内容能够直接触发学生的评论冲动和表达热情，根据学生的反馈对教学内容做出修改、补充或更新，就实现了学习者与学习内容的交互。

为了建设“交互型”学习资源，需要首先实现知识点的结构化管理、组织和跳转。在知识库技术和语义标注技术的支持下，通过改变传统纸质教材的线性结构表现方式、从教材文本中自动识别出包含的知识特征、并根据教材知识本体和教学论自动标注学习内容，能够建立全新的知识组织形式，从而根据学习的目标和学习内容自动生成学习计划，实现个性化的学习内容[375]。

增强现实技术被应用于辅助对外汉语立体化教材建设。立体化教材是指依托现代教育技术，以能力培养为目标，以纸质教材为基础，以多媒介、多形态、多用途及多层次的教学资源和多种教学服务为内容的结构性配套教学出版物的集合。焦燕[376]利用 HP Reveal 平台，采用手持式增强现实设备和基于人工标记的技术对《汉语教程》第一册（下）第 26 课《我打算请老师教京剧》进行了立体化教材制作尝试，将对“语法”的讲解录制成视频，选取了练习中的一段会话练习制作为动画，对生词“让”进行录音，运用 HP Reveal 平台实现增强现实功能。

利用多项数字化交互技术实现的智能化教材是数字教材发展的新阶段，融教学内容、学习工具和个性化学习服务于一体，全方位支持学习者与教材之间的互动。目前主要有《七色龙》（外语教学与研究出版社）、《中文听说读写》（美国 Cheng & Tsui 出版社）、《牧羊犬“丁丁”》（美国 iChineseReader 中文学习平台）等。

以《七色龙》智能教材为例，《七色龙》是面向海外主流中小学及国际学校 K-6 阶段的学习者开发的模块化中文教学资源，共包含 15 个主题、3 个难度级别、225 本读物。全套资源包含分级读物、配套学习 APP、教师资源、评估测试方案等产品形式，拥有一体化的产品体系，既可以满足课堂教学的需求，也可以满足学习者课外强化和拓展的需求。

9.1.3 数据资源库建设

1. 资源及标准化建设

国际中文教育等级标准的建设经历了 30 多年的探索，2021 年《国际中文教育中文水平等级标准》（以下简称《等级标准》）的发布与实施是国际中文教育学科与事业进一步走向标准化、规范化、国际化的突破性创新和原创标志性成果[377]。

《等级标准》建设由“二维基准”发展到“四维基准”，目前实现对音节、汉字、词汇、语法四大要素的标准涵盖。突破之一是将汉语音节作为突破口和创新点融入新模式，具有鲜明的汉语特色；二是创新性融入“语法标准”，提炼出三等九级共 572 个语法点。《等级标准》的“语言量化指标”中规定了达到每一级汉语水平应掌握的音节、汉字、词汇、语法四个维度的汉语基本要素的数量，

适用于国际中文教育的学习、教学、测试、评估等各个方面。

2. 汉语教材库

汉语教材库是汉语教材资源的集合，是大数据下的数据集成。现有中山大学全球汉语教材库，收录对外汉语教材 16000 余册（种），

其中有10000余册（种）实体教材，目前，全球汉语教材库可支持4种功能：检索教材、样课预览、查询教材、上传共建。另有全球华文教材语料库，规模约200万字；澜科语言科技中心构建的汉语教材语料库等。

3. 文化教学资源库

文化教学资源库主要以网站形式呈现。从内容上看，涉及当代中国社会（基本国情、旅游地理、节日习俗、饮食、休闲娱乐、日常生活、中外对比）、中华传统文化（历史、文学、艺术、思想）、语言文化要素（汉字、词汇）等。除此以外，还有针对古诗词资源、博物馆资源等进行的文化资源库的开发建设。另外，也有一部分地方文化资源，包括西北少数民族文化资源、广西边疆少数民族资源、齐鲁文化资源、辽南文化资源等，各地文化资源各具特色。

4. 国际中文教育传播资源库

关于国际中文教育传播的相关资源库有北京语言大学“汉语国际传播动态数据库”，目前已经完成了数据库总体架构及字段设计，共设6个一级子库，36个二级子库，下设698个一级字段，收录各类数据共计13125条，案例132个；北京外国语大学“中华文化走出去动态数据库”，包含中华文化走出去的相关数据，主要包括出版、文化艺术、传媒、影视等相关信息；暨南大学“华侨华人特色数据库”，包括华侨华人书刊全文数据库、华侨华人学术资源数据库、华侨华人政策法规数据库等子数据库。中央民族大学“国际汉语教学数据库”，包含孔子学院数据库、教学案例库、国际汉语教材库、中华文化国际传播课件库、汉语国际教育硕博论文库、期刊论文库等子库。

9.2 智能技术赋能教学实践

智能技术从算法发展和技术应用角度划分，大致经历了程序模型、概率模型和深度模型三个阶段。在程序模型和概率模型阶段，人工智

能以计算机辅助教学、计算机辅助学习等形式服务于教育行业，以程序化处理、结果反馈以及简单推理等为特征。进入深度模型阶段后，随着算法模型的改进和计算能力的突破，人工智能在系统化、智能化方面大大增强，能够胜任复杂推理任务，其在教育行业的应用不断深化，以 AI 互动课程、个性化学习、人机互动和智慧教育等为典型应用。人工智能正在改变教育行业，为教育发展提供动力，减轻教师负担，提升学习效果，提高教育教学的质量和效率。

9.2.1 虚拟现实与沉浸式教学

虚拟现实技术（Virtual Reality，简称 VR）是一种先进的人机计算机接口技术，它利用计算机生成一种高度逼真的、模拟人在现实环境中进行视、听、动等行为的虚拟环境，并通过多种传感设备，使人投入到该环境中，实现人与该环境间的自然交互。基于虚拟现实的沉浸式教学一般需要多项智能技术的综合，包括：

（1）智能系统。包括语音管道和手势识别。语音管道记录和转录学生的话语，并对从转录文本中检测到的意图进行标记。由骨骼跟踪设备和自定义手势识别软件启用的手势流，提供有关用户做出哪些手势的输入。

（2）多模式推理。个体模态交互包括话语的音调分析和手势识别。从组合模态推断的交互包括解释指示性话语，并结合指向手势识别意图。

（3）多模式演示。合成语音、环境音频、特效和沉浸式游戏视觉效果在前端呈现系统响应，以完成多模式通信循环。

国外 VR 语言学习项目发展迅速，如伦斯勒普通话项目这样较大规模的项目已经开始使用如 360°全景屏幕、无标记运动跟踪传感器阵列等先进技术[378]。

国内很早就开始了结合虚拟现实技术和国际中文教学的尝试。周晓军&马君[379]基于 VRML 技术，综合多媒体技术，设计了情景模拟

教学。此后马君[380]又实现了基于 VRML 的远程对外儿童汉语教学课件设计。但受限于当时的软硬件条件，未能投入实际应用。刘哲[381]借助 VRChat 平台，通过将 VRChat 与体演文化教学法相结合，进行了一次较完整的教学实践案例，总结应用 VRChat 进行汉语体演文化教学的教学效果、注意事项、现有不足，为 VR 沉浸式汉语体演文化教学设计开发提供参考和范例。但相比国外，在硬件条件、技术水平、项目规模方面，都仍有一定的差距。

增强现实（Augmented Reality，简称 AR）技术是在虚拟现实的基础上发展起来的技术，是指通过将计算机生成的虚拟场景、文字注释等信息实时、精确地叠加到使用者所观察到的真实世界景象中，对人的视觉系统进行延伸和扩充。增强现实技术具有虚实结合的特性，契合了当前第二语言习得理论强调本地化、上下文学习和与现实世界的有意义联系的新思想。增强现实技术能够为学习者提供各种拟真的认知场景，提高学习效率，为学习者提供个性学习的发挥空间；在增强现实技术搭建的学习场景中，学习者不仅可以同其中的学习对象互动，也可以同其他学习者实时互动，交流经验。所以，从理论上来看，增强现实技术用于辅助学习具有很大的潜力。增强现实系统分为基于位置的系统和基于图像/对象的系统[382]。目前教育环境使用最多的是基于图像/对象的系统。

温韫[383]利用基于增强现实的汉字组合游戏辅助小学生协作汉字学习，表明了引入增强现实技术能有效提高初学者（特别是汉语水平较低的学生）的汉字拼写知识学习。张胜兰[384]通过为期三周的课程，实践了将增强现实融入基于任务的主题语言教学单元，通过与学校商店的合作，引导学生探索发现可供交互的神秘商品和任务，教授与购物、服装、色彩等相关的中文词汇和句子结构。Sinyagovskaya, Daria 等[385]则基于增强现实技术开发了一款发音训练应用程序。

虚拟现实技术用于国际中文教学仍处在探索阶段，目前的尝试都

存在各自的缺点和不足，受限于技术开发能力，所能提供的沉浸式情境也较为有限。希望在今后的研究中，能够创设更加丰富、多元的情境，并以此为依托，研发出一系列适合国际中文学习者使用的结合虚拟现实技术的应用软件。

9.2.2 中文学习智能纠偏

1. 口语发音纠偏

口语发音作为语言学习中的一个重要环节，在国际中文教育中面临着学习者“中文难”的心理障碍问题。其中口语部分尤其难在声调，在没有环境条件的情况下，难以实时指出和发现读音中哪个音标、音调、声韵母读错、误读等情况，不能发现字词句篇章哪一句话读的标准。通过 AI 技术，能够诊断声韵调等典型错误，纠正发音问题。这涉及到语音预处理、评测声学模型自适应、评测特征提取及评分映射等多个环节。

计算机辅助发音训练系统的核心模块主要有发音自动评价和发音偏误检测。发音自动评价指对发音人的发音进行正面打分，适合评估学习者的整体发音水平；发音偏误检测识别学习者的错误发音，并给出改进建议，对学习者在之后的学习中改善错误发音有积极的影响。

目前主流的发音偏误检测系统都是基于自动语音识别的框架。深度神经网络近些年在自动语音识别应用中取得了较大的成功，显著降低了语音识别错误率，相比高斯混合模型，深度神经网络采用拼接帧作为输入，同时具有深层结构，比浅层结构的高斯混合模型具有更强的模型表达能力。如张劲松等[386]应用深度神经网络进行声学建模,比较 Mel 频率倒谱系数、感知线性预测分析系数和 Mel 滤波器组系数 3 种声学特征参数,并利用网格联合技术整合 3 种声学特征得到候选网格，进而实现对语音的表达。

但全连接深度神经网络参数多，需要大量样本进行训练。带标注的发音偏误样本过少容易引起深度神经网络过拟合，因此，有学者尝

试通过卷积神经网络来解决这些问题。如甘振业等[387]利用深度全序列卷积神经网络和链接时序分类技术,建立了一种用于发音偏误检测和诊断任务的端到端语音识别方法；杨龙飞等[388]应用卷积神经网络进行声学建模，通过实验证明卷积神经网络比之全连接深度神经网络检测正确率相当，虽有稍高的错误拒绝率，但是获得了更低的错误接受率。

工程应用方面，科大讯飞开发了 FiF 评分系统，实现了产业化应用。该系统共包含 3 个模型：（1）语音识别模型，用于识别被试的话语；（2）标准发音模型，用于判断发音准确度；（3）通用分数映射模型，通过收集大量按照题型区分的口语测试数据提取评分维度特征，并聘请专家对口试录音进行评分，基于 SVM（Support Vector Machine）分类器和非线性回归映射算法，实现维度特征到人工评分的高精度映射（包括特征到单项分的映射）。该系统可以从发音准确度、重音、流利度、内容完整度四个维度给学生的口语表现打分，每个维度又包含若干所提取的评分特征。

2. 汉字书写纠偏

计算机辅助汉字书写教学技术的任务是借助各种数字手写设备，综合利用汉字信息处理、计算机图形学、数字图像处理、人工智能、文字学等领域的相关知识，研究汉字书写规范性的智能化、自动化评判方法以及可视化的用户反馈形式。它的关注点在于评判内容（各种书写错误及书写规范）的全面性和准确性，反馈效果的直观性和启发性。它的最终目标是实现学习者在无人值守的情况下进行汉字书写的自由练习。

在智能技术的协助下，目前计算机辅助汉字书写教学已从初期的单向数字化演示逐渐转向汉字书写规范性的智能评判，即标明用户在书写过程中的错误和缺陷并予以纠正反馈。其关键技术环节包括字形匹配和反馈指导。

（1）字形匹配

字形匹配是指建立手写字与模板字之间的笔画对应关系。近年来，针对字形匹配技术的研究有很多成果。例如，Hu 等[389]首先将汉字的笔画位置关系表示为属性关系图（Attributed Relational Graph），然后通过将笔段投射到坐标轴上实现书写信息的裁剪，最终建立起模板字和手写字之间的匹配关系；Chen 等[390]根据斜率将手写字的笔画进行归类，然后与模板汉字进行匹配；Tang 等[391]使用动态规划算法进行字形匹配；吕晓晨等[392]提出了一种针对脱机手写字图像的字形匹配方法；安维华等[393]提出了一种基于最优化模型的联机手写汉字字形匹配方法；吴嘉伟[394]提出了基于松弛匹配的字形匹配算法：首先定义笔段之间的相似性和笔段之间的相容度，然后利用笔段之间的相容度对笔段相似性进行迭代调整，便可得到最大化的匹配结果。

（2）反馈指导

如何从适当的粒度（笔画、部件、整字）出发进行错误反馈和书写指导，是智能汉字书写纠偏仍有待解决的课题。马乐慧[395]提出了一种基于字形相对中心的事后评判算法，通过对各种笔画参数的差异性进行聚类分析，达到了定位手写汉字中关键书写缺陷的目的，在一定程度上实现了无人值守的评判目标。

未来的计算机辅助汉字书写教学技术，将以汉字书写规范性评判和水平评测为研究重点，以全面化、精确化和智能化为主要目标，并且拓展更多的应用场景。

9.2.3 基于知识图谱的个性化学习

知识图谱主要技术包括知识获取、知识表示、知识存储、知识建模、知识融合、知识理解、知识运维等七个方面。通过这些技术的综合运用，能够帮助实现学习者个性化学习。

由于学习者的中文学习水平和能力参差不齐，如何让老师了解每个学习者的学习水平，进而定制化或个性化的推荐相应的学习资源是

非常困难的。传统教学模式下，老师给所有学生布置作业都是统一的，无法很好地提供针对性的学习辅导。利用人工智能和大数据技术，提供智能化教学解决方案，能够实现个性化语言学习，辅助老师全面了解每个学生的中文学习现状，便于提供更精准的教学指导，提升教学效率。其主要过程包括：

1）对学习者的学科能力维度、主题语境维度、书面表达维度进行用户画像建立，更好的掌握学习者情况；

2）基于汉语语言学习能力的学科知识点分析方法，通过对学生的空间想象能力、抽象概括能力、推理论证能力、运算求解能力、数据处理能力的综合分析，全面诊断学生的学业水平，实现语言能力分析；

3）结合大数据等统计学方法，在学习者学习、测验过程中预设数据采集点，通过听说读写等几个维度反复练习，跟进判断学习者水平等级；根据用户使用的资源和产生的数据，根据用户使用习惯和学习路径，进而进行数据标注，形成因子图，产生学习和知识的推理；

4）根据教学知识图谱，进行学习路径拓展，进而对其进行个性化资源推荐。

例如，科大讯飞通过针对学习者建立基于 HSK/YCT 标准的、细化到四级的知识点体系，完成了习得顺序的标注和建立，同时完成包括海量题库、课本章节、名师微课等资源与知识点的关联工作，并将人工专家标注数据与自然语言处理、数据挖掘分析等技术相结合，研发训练了资源到知识点的自动关联算法，为新资源的持续积累和建设提供了很好的技术支持；基于语文教学的知识点能力分析模型，实现了在多次练习数据分析的基础上，掌握学生薄弱知识点，形成自动智能分析判断，形成个性化学习资源的推荐。

9.2.4 移动学习

随着移动通信技术的发展，基于移动计算技术的应用也越来越丰

富，其在教育领域中的应用带来了一种崭新的学习方式——移动学习。移动学习是基于无线移动通信设备来获取教育信息、资源和服务的一种学习形式。在无线网络技术和移动设备的支持下，学习者能够借助移动终端从大量的学习资源中获取所需知识和教育服务，也是新媒体技术与教学活动融合下具有双向交流和交互功能的一种新型学习方式。

即时通讯工具凭借其发送内容的丰富性、互动交流的迅捷性、链接材料的方便性等诸多优势，迅速赢得了人们的青睐，也成为在华留学生交流沟通方式的首选。将即时通讯工具的这些优势运用于国际中文教学，可以促进留学生语言能力的发展。

随着计算机网络技术的发展，尤其是随着 Web2.0 技术的发展和普及，社交媒体等新兴平台形式极大地改变了人际交流与沟通方式。国际中文的基本教学原则要求学生掌握汉语的基础知识和基本技能，培养运用汉语进行交际的能力。随着社交媒体在人际交流与沟通中的广泛应用，对于中文学习者来说，培养自己的言语交际能力离不开与我们生活相关的社交媒体。因此，国际中文教育界开展了一系列基于任务法和交际法的社交媒体教学尝试。

理论研究方面，徐品香[396]基于社会文化理论和活动理论提出了 IAST-A 模型。其中，IAST 分别代表互动（Interaction）、音（视）频（Audio-video）、分享（Share)及文本(Text)，最后一个 A 则代表活动（Activities）。徐品香将教学活动分成三个层次：中心层是面对面班级教学，侧重点在基础知识和基本技能的训练与掌握；中间层是社会性网络与应用；外层则是这些社会性网络支持的具体教学活动。

不少学者对社交媒体教学的效果进行了评估。陈珂忆和辜亿珈[397]分析了中文社交媒体上对外汉语修辞教学资源现状，指出社交媒体上的汉语学习资源以基础语音、简单汉字及生活词汇、日常文化知识教学为主，缺少高水平的修辞教学资源，难以替代传统的修辞教学

课堂。李代鹏[398]发现，基于社交网络的对外汉语教学无法提升学习者整体的汉语能力，但对于提升学习者的汉语词汇、阅读和写作能力具有显著作用。谢静[399]则通过对教师的调查发现，社交媒体可缓解师生交流的压力感和紧迫感，有助于提高留学生口语和书面语输出的正式度和正确率，也有助于教师根据学生特点开展个性化教学。但社交媒体上的网络资源存在知识内容分散、碎片化严重、语言素材良莠不齐等问题，社交媒体本身也容易干扰、分散学生注意力。

实证研究方面，潘文斌[400]将任务型教学法融入到社交网络平台（Pinterest）之中，以检验社交网络平台对于任务型写作教学的实践功能；刘丽莎[401]探索了将 Facebook 应用于汉语文化教学的可能性及具体方式；Veronika Seroshtan[402]尝试通过 Instagram 上的汉语教学进行视觉辅助工具的设计与应用。

9.2.5 智能技术辅助教学分析

教学分析以教学过程分析和教学切片分析为研究对象，通过对教学事件及相关因素的计算与分析，能够监测教学并精准定位，从而使教师可以采取科学高效的教学干预策略，为教学过程的绩效管理提供保障[403]。教学分析的对象是复杂的，这要求从多个维度空间把握研究内容，既要考虑教学的结构构成和过程实现，又要考虑教师、学生、教学环境及其一切交互表现形式和结果。以往由于研究手段或技术工具的限制，研究者主要关注某些特征或关系，其研究和描写多为断点式或结论性的，由此得到的思辨或实证研究结果常常是对教学统中某个点或某个侧面的画像，未能反映复杂教学系统的全貌。

以大数据技术为代表，智能技术为记录和描写课堂教学提供了条件，使得教学系统运行过程中产生的海量数据得以保留，这些信息记录了教学发生、发展及变化的全过程。对这些数据进行挖掘和利用，所得的量化研究结果为创新国际中文教学带来了强大的驱动力。

智能技术辅助下的教学分析需要经历如下几个阶段：对研究问题

进行分析，数据采集和数据诊断，数据特征收集和模型发现，对特征或模型进行分析和解释。

1）根据研究问题确定相关要素:即研究问题映射到数据结构的过程。对要素及其结构的设置，决定了最终可以对数据进行怎样的分析以及可能得到什么结果。

2）数据采集：用于研究的数据可以是一切来源于教学的原始数据。可以从已有的包含数据结构的数据库中提取相关数据；也可以从不包含数据结构的数据库中按照属性结构对应的数据结构去采集数据；或者重新收集语料建立新的数据库。

3）数据诊断：初步判断数据整体质量后发现和处理异常值。聚类方法、分类方法和回归方法通常被用于数据诊断。各种数据可视化手段，如绘制或生成折线图、散点图、柱形图等，也可用于甄别出异常的数据点。

4）特征收集和模型发现：通过聚类、分类、或两种方法联合使用，挖掘教学过程中各要素间的关联关系、因果关系和序列模式。

5）对特征或模型进行分析和解释。

目前，郑艳群等已通过智能技术辅助下的教学分析和教学计算，对汉语阅读教学[404]、听力教学[405]、口语教学[406]、写作教学[407]、综合课教学[408]等课程的教学结构进行了分析，并针对教学过程建立了理论模型和应用模型。

未来，智能技术辅助下的教学分析和教学计算可以将教师的教学行为进行量化，提高教学质量评测的精准度和客观性，使个性化且全面的教学质量评价与反馈落到实处。

9.3 智能技术赋能中文测试

9.3.1 智能中文测试

语言测试是测量学习者语言能力、评估语言教学效果的重要手段。

长期以来，节省编制题库所需的大量人力和克服经典测试理论的缺陷始终是研究者感兴趣的课题。智能技术在这两方面都能够提供极大的帮助。

1. 自动化项目生成技术

自动化项目生成是随着计算机技术发展而逐渐兴起的，它是指根据开发者要求，在项目生成算法的指导下，自动地生成符合参数的项目。早期的自动项目生成主要采取项目模型法，即指将经过检验且指标良好的项目作为模板，通过改变和替换与问题解决难度无关的描述，组合形成多个新项目的过程。

国际中文测试领域很早就开始尝试自动化项目生成技术。北京语言大学早在 1998 年就研制了“中国汉语水平考试（HSK）题库试卷自动生成系统”，该系统能按照一定的命题计划生成 HSK 试卷，后续也开展了一系列有关自动生成试卷的原则及参数、统计分析、与人工命题试卷对比等研究[409]。

近年来随着人工智能技术的广泛应用，自动化项目生成的算法也有了新的进展，语义分析和深度学习模型都已被用于自动化项目生成。国际中文测试也正在逐步尝试这些新技术。

2. 计算机自适应测试技术

传统的纸笔测试基于经典测量理论，所有的考生无论其语言水平差异有多大，都必须在相同的时间内完成由相同题目构成的定长测验。语言水平较高的考生在回答难度较低的部分题目时无法获得有效的分数差异、而语言水平较低的考生面对难度较高的部分题目时则无法提供有效的测量信息；同时，答对相同题目数量的考生被视为具有同等语言能力，这忽视了题目难易的差别。而基于项目反应理论、借助计算机技术和网络技术实现的计算机自适应测试，则能够克服上述缺点，从而达到更高的信度。

北京语言大学谢小庆教授等从 2003 年开始研究计算机化 HSK 自

适应性考试，并开发了模拟 HSK 考试系统和练习系统。谢小庆[410]讨论了采用计算机自适应测试技术后，不同难度题目测试分数的等值问题，提出了共同组等值、共同题等值和分半组合等值三种方法。柴省三[411]从理论上解释了计算机自适应测试的原理，并设计了计算机自适应测试逻辑过程。

3. 自动化项目生成技术和计算机自适应性测试技术的结合运用

胡韧奋等[412]尝试了同时运用自动化项目生成技术和计算机自适应性测试技术构建词汇考试系统。通过使用多种自然语言处理（NLP）方法从大规模语言资源中自动提取属性值构建词汇知识库，制定了选词题、发音题和搭配题的具体生成流程，最后将生成的题库用于计算机自适应性测试实验。该项目表明这两项技术的结合可以有效地构建测试项目并显著降低测试成本。此外，计算机自适应性测试的测试结果可以为自动化项目生成算法提供有价值的反馈。

尽管国际中文教育在智能测试方面起步较早，但总体而言进展缓慢，目前的研究局限于理论研究和模拟测试，缺少能够落地的真实应用场景。如何将理论探索转化为工程实践仍有待进一步努力，智能测试的深入研究也需要更多的实证研究与真实测试场景提供支持。相信在未来，通过获取真实教学大数据、结合《国际中文教育中文水平等级标准》等教学大纲提出的知识内容和认知能力目标，综合运用多项智能技术的智能测试系统将有能力最大程度地自动化生成适合真实教学环境与测试场景的高质量评测项目。

9.3.2 中文句法错误自动诊断

近年来面向英语学习者的作文自动批改技术发展迅速，成为了语言信息处理领域应用研究的新热点，也引起了国际中文教育界的关注，并尝试开展面向汉语作为第二语言学习者的中文句法错误自动诊断。早期的中文句法错误自动诊断一般运用统计机器学习技术、规则分析方法或将两者结合。在引入深度神经网络方法后，中文句法错误自动

诊断获得了快速发展。由于其效果要明显好于传统的统计建模方法，当前几乎所有的中文句法错误自动诊断研究都选择了基于深度神经网络的方法。

最先被采用的是 CNN-LSTM 技术。输入句子表示为单词序列并生成句子矩阵，并在句子矩阵上使用单个卷积层来提取特征。通过在整个矩阵上滑动过滤器来获得完整的卷积，每个过滤器对句子矩阵进行卷积操作并生成特征图，然后使用池化层对每个特征图上的特征进行二次采样。LSTM 用于向量组合的顺序层，在 LSTM 记忆单元顺序遍历所有特征向量后，将顺序层的最后一个状态作为神经计算的输入。最终的 Softmax 层接收计算结果并使用它对句子进行分类。

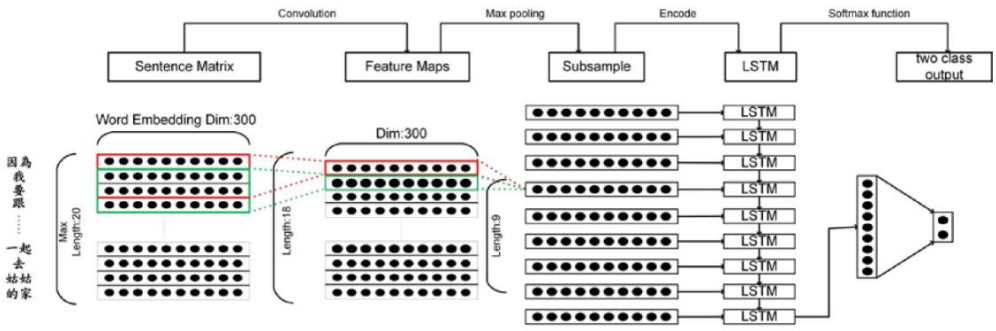


图 9-1：用于中文语法错误检测的 CNN-LSTM 模型示意图[1]

在 CNN-LSTM 的基础上，策略梯度 LSTM 模型、BiLSTM-CRF 模型等技术纷纷被用于中文句法错误自动诊断，不同程度地提高了中文句法错误自动诊断的准确率和召回率。

近期，Transformer-based network architectures（如 BERT，RoBERTa，XLNe，ELECTRA）在很多自然语言处理任务中取得了良好的表现，这一技术也被引入了中文句法错误自动诊断。

双向编码表示转换模型（Bidirectional Encoder Representations from Transformers，BERT）是基于 Transformer 架构的预训练语言模

型，解决了经典神经网络的数据依赖和语言编码问题。李琳、董璐璐、马洪超[413]基于 BERT 进行了实验研究，发现池化策略对模型性能有显著提高，抽取某个编码层进行池化的效果要好于多个编码层拼接在一起进行池化的效果。

ELECTRA 是一种基于对抗性学习的深度神经网络模型，李龙豪等将其用于中文句法错误自动诊断[414]，在实验中取得了很好的效果。ELECTRA 训练两个转换器模型：生成器，它替换序列中的标记以训练掩码语言模型；鉴别器，它试图识别序列中的哪些标记被生成器替换。

深度神经网络方法大大促进了中文句法错误自动诊断的发展，但中文句法错误自动诊断仍面临着缺乏足够语料的困难。目前中文句法错误自动诊断最主要的语料来源是北京语言大学所构建的 HSK 动态作文语料库与台湾师范大学所构建的 TOCFL 华语文作文语料库，所能提供的语料数量较为有限且增长缓慢，难以支撑深度神经网络模型对训练数据的规模要求。同时，国际中文教育实践中产生的汉语学习者作文语料往往缺少准确的错误标注，数据质量的参差不齐也影响到了深度神经网络训练的效果。为此，已有研究[415]尝试通过基于简单文本增强法（EDA）的数据增强方法自动合成语法偏误数据集，取得了一定的效果。

中文句法错误自动诊断的进一步发展是主观题中文作文批改技术。主观题中文作文批改技术提供包括异常检测、多维度批改、总评与分项评语等一体化的语文作文自动评阅解决方案，还包括针对诸如文本通顺、文采、立意分析、篇章结构等难度较高的维度进行探索。功能包括：语法错误诊断，以预训练语言模型为基础，结合少量标注数据和大规模自动构建的伪数据，进行错别字、语法以及标点、成语等多类型错误识别；篇章结构质量评估，通过识别句子和段落的论辩角色（如引论、主旨、论点、论据等）来表示篇章结构，提出了层次

多任务学习方法，融合句子级和段落级的篇章角色表示进行议论文篇章结构质量评价；优美表达识别，利用众包手段从多个来源采集人们推荐的优美句子表达、描写表达等，构建识别模型。不同层面的深度语言分析拓展了传统作文自动评分系统考察的评分维度，构建相应特征有助于提高评分模型的鉴赏判别能力和评分准确性，并为评分提供了更好的可解释性，减轻阅卷人力、财力负担，促进自动批改在课堂教学场景落地，辅助教师课堂教学。

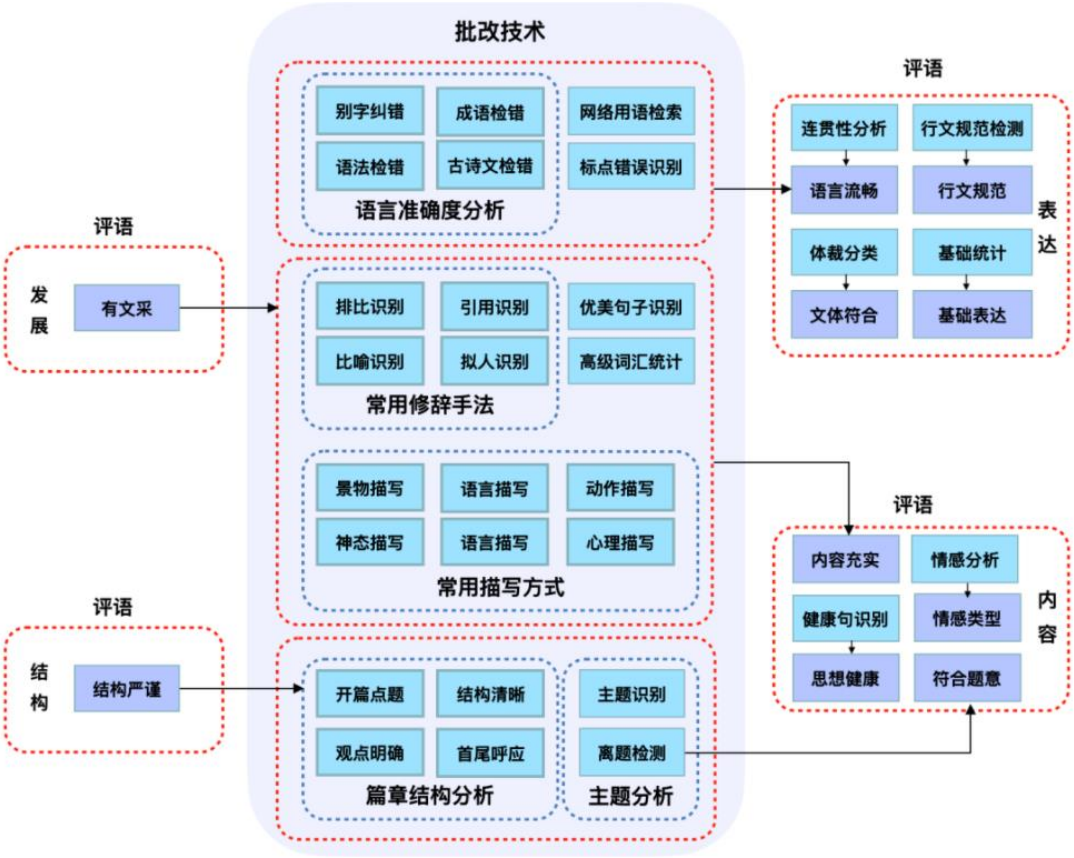


图 9-2：科大讯飞主观题中文作文批改技术框架图

9.4 语料库技术辅助国际中文教学

作为语言智能处理的有效方式和手段，语料库技术在国际中文教育领域发挥的作用越来越明显。语料库以科学抽样和大规模加工为主要特点，集合一定数量的语言材料，实现对真实语料的储存、检索和

提取，能够为语言学和国际中文教育提供智能处理的新路径。当前，国际中文教育领域的语料库建设和发展不断推进，基于语料库的相关教学研究和成果丰硕，全球汉语中介语语料库等多个具有专门性和针对性的语料库相继建成，共同助力国际中文教育智能处理取得新发展。

9.4.1 基于语料库的国际中文教学研究

运用语料库技术，国际中文教学相关研究能够将定性与定量的方法相结合，针对汉语作为第二语言学习者的真实语料进行描写和分析，从而使研究结论具有较强的客观性、普遍性和稳定性，提高汉语作为第二语言教学研究的水平。近年来，基于语料库的国际中文教学研究数量较多，范围较广，以下列举其中三类：

第一，具体词语或词类习得研究。这类研究选取具体的词语或词类作为研究对象，通过语料库检索，观察研究对象在真实中介语语料中的分布情况，进而进行描写和分析。相关研究有：夏历等[416]通过暨南大学留学生书面语语料库和 HSK 动态作文语料库，对留学生汉语列举类词语“例如”和“比如”的使用情况进行描写。李治平等[417]基于 HSK 动态作文语料库，对留学生汉语语篇关联语使用情况进行统计和分析。周睿[418]运用 HSK 动态作文语料库和日语学习者书面语语料库，对汉语和日语中的强程度副词的二语习得情况进行对比。

第二，词汇习得与发展研究。这类研究基于语料库技术，从更加宏观的视角出发，考察二语学习者的词汇习得及习得发展情况。相关研究有：张江丽[419]自建外国留学生汉语笔语语料库，考察汉语二语者与汉语母语者在产出性词汇量上的差异。林柱等[420]基于汉语母语者 BCC 语料库，概括出汉语集合量词使用规则，进而考查集合量词习得顺序。刘竹林[421]将 85 组误用词和当用词放入 800 万字的汉语中介语语料库中进行检索，考察学习者词语跨类混淆的分布特征。

第三，句式及句法习得研究。这类研究通过在具有句式和句法标注的语料库中进行检索，以描写汉语学习者对于汉语典型句式或句法

结构的习得情况。这类研究要求对语料进行相关标注，研究数量相对较少，典型的有：谭晓平[422]对 HSK 动态作文语料库中韩国汉语学习者的“比”字句进行穷尽性的检索，探究“比”字句相关结构的习得情况。郝瑜鑫等[423]基于自建句法标注中介语语料库，对英语母语背景学习者汉语动词配价发展进行了计量研究。

9.4.2 国际中文教育语料库资源建设及标准探讨

随着语料库技术在国际中文教育领域的逐渐运用，基于语料库的资源建设不断取得新的成果，相关语料库资源及技术应用不断进行，并不乏对语料库的建设规范和原则等进行讨论。

第一，语料库资源建设。语言本身就是一种资源，国际中文教育相关语料库的建设，亦是该领域的资源建设。刘华等[424]基于现有中医汉语类教材、中医专业类教材、中医网站三大语料来源，建设中医汉语语料库。周小兵等[425]讨论国际汉语教材语料库的建设和应用，指出教材语料库在教材编写指南的研制、教材评估与难度测定、测评软件的研制与使用等方面的积极价值。王敬等[426]选取 1181 个重点多义词，在 197 册经典汉语二语教材上进行了多义词词义标注，构建规模约 350 万字的面向汉语二语教学领域的词义标注语料库。

第二，语料库技术应用。基于语料库相关技术的词汇提取、词表制定等，在国际中文教育领域亦有广泛应用，相关成果如：王治敏等[427]利用大规模语料库重点考察词语在历时文本中的分布特征，并建立可反映动词历时变化的汉语常用词语统计词表，为对外汉语教学提供参考依据。张引兵等[428]基于所给定的语料，运用词汇构词知识库和类推机制，给出了词汇对相应语料的综合覆盖贡献度评价方案，为今后各类词汇大纲的制定及完善提供思路和方法上的参考。王贵荣等[429]从 BCC 语料库中抽取动宾搭配知识，并对抽取结果进行了初步消歧，最终获得动宾搭配 300 万对，形成动宾搭配知识库。

第三，语料库建设标准和原则探讨。在语料库工程建设过程中，

也有学者不断反思，对语料库建设标准和原则等提出探讨。张宝林等[430]认为，语料收集应遵循真实性与代表性、平衡性与系统性、有声性与有图像性的标准；语料标注应遵循全面性与相对性、科学性与通用性、只标不改等原则。黄友[431]提出，面向二语学习者的汉语易混淆词语词典和语料库建设应遵循以下原则：应编纂针对不同母语学习者的汉语易混淆词语词典；词典收词以易混淆词语为对象；释义元语言应简明易懂；词典编排应突破传统程式，采用“先给例句，然后根据是否可替换来给出思考提示，最后描述区别”的编排程式；词典编纂应依托语料库。胡晓清[432]指出，建立国别化汉语中介语动态语料库是对原有通用型汉语中介语语料库的重要补充，是汉语中介语的两翼之一，并提出立体化建库理念将取代平面化建库理念的论断。

9.4.3 国际中文教育代表性语料库实例

基于语言智能处理和语料库技术，国际中文教育领域相继建立众多实用语料库。以下选取其中具有代表性的 5 个语料库进行介绍，其中前三个为专业型语料库，后两个为通用型语料库：

（1）全球汉语中介语语料库，是因应汉语作为第二语言教学的学科建设和科学研究的需要而设计建设的一个迄今为止规模最大的汉语中介语语料库，在设计理念、建设策略与方式、标注内容与方法、数据统计、检索方式等方面具有首创性，是语料库建设 2.0 时代具有代表性的语料库。目前收入原始语料约 2367 万字，从汉字、词汇、句子等 10 个层面进行了标注[433]。该语料库面向海内外用户免费使用，网址为：qqk.blcu.edu.cn。

（2）HSK 动态作文语料库，是母语非汉语的外国人参加高等汉语水平考试（HSK 高等）作文考试的答卷语料库，收集了 1992-2005 年的部分外国考生的作文答卷，语料总数达到 11569 篇，共计 424 万字。该库自 2006 年上线以来，为对外汉语教学与相关研究提供了量化分析的坚实基础，推动了基于语料库的汉语教学研究、习得研究和

中介语分析的发展，取得了很好的学术效益[434]。网址为hsk.blcu.edu.cn。

（3）汉语阅读分级指难针，旨在为国际汉语教师提供阅读文本的难度定级与智能改编，共包含“文本定级”“词语标注”和“字词档案”三个核心模块。其中“文本定级”工具以权威词汇等级大纲为参考标准，通过算法生成文本难度值，为文本难度提供数值结果[435]。网址为languagedata.net/editor。

（4）北京大学 CCL 语料库，是面向语言学本体研究和语言教学的大规模语料库，目前包括现代汉语、古代汉语和汉英句对齐平行语料，规模超过 7 亿汉字。CCL 语料库检索系统以包括汉字、字母、标点等在内的字符为基本索引单位，提供普通查询、批量查询、模式查询等多种检索方式[436]。该语料库在海内外汉语研究和教学领域得到了广泛应用，产生了较大的影响。

（5）北京语言大学 BCC 语料库，是以汉语为主、兼有其他语种的在线语料库。该库总规模达数百亿字，是服务语言本体研究和语言应用研究的在线大数据系统。该库为语言本体研究提供数据和技术支持，在大数据背景下，可以证实、证伪或者发现语言现象；作为语言应用开发的基础平台，为信息抽取、构建知识图谱、语言自动分析等提供便利；同时，也为语言教学研究提供统计数据和实例支撑等[437]。

9.5 基于智能技术的国际中文综合教学平台

随着中国综合国力的不断提升，国际中文教育事业近年来取得了长足的进步与发展，全球范围内的中文学习热潮也不断涌现。此外，从 2020 年开始爆发的“新冠肺炎”疫情给国际中文教学事业带来挑战的同时，也深刻地改变了其未来的发展方向，基于互联网的“线上”教学在未来一段时间内仍将是主要的授课方式之一。不同于传统的国际中文“线下”课堂教学，“线上”中文教学是借助互联网，进一步

发挥多媒体技术、虚拟现实技术、人工智能技术、自然语言处理技术和大数据技术的优势，衍生出的一种全新教学形式。

正是在这样的背景下，各种中文教学 APP 和国际中文教育教学平台快速发展，极大地满足了国际中文线上教学的需求。

9.5.1 中文教学 APP

教学应用资源主要指国际中文教育类 APP，目前在应用市场中的中文教学 APP 数量很多，但也存在一些 APP 用户使用量很低和已经下架的 APP。现有的汉语学习 APP 从内容上看，可分为语言要素类、语言技能类、专项内容类、专项功能类和其他五大类。其中，语言要素类 APP 又分为语音、词汇、汉字类，词汇和汉字学习的 APP 较为丰富，且常用的 APP 目前尚未出现语法专项训练类 APP。语言技能 APP 以语言综合训练 APP 为主，听说训练次之。网络技术的发展也使得 APP 的研发逐渐走向智能阶段。语音识别技术、文字识别技术、深度学习技术等在中文学习 APP 中实现越来越丰富的应用。

当前用户数量丰富、用户推荐较多的综合类汉语学习类 APP 主要有 Chinese skill、Hello Chinese、Super Chinese、SPK Chinese 等；汉语要素类 APP 主要有 Pleco、Skritter、Pinyin News、嗨中文等。从语种支持上看，当前汉语学习 APP 大多支持中文简体和英文，少数 APP 支持多种语言。

表 9-1：部分汉语学习 APP 基本信息介绍

中文学习 APP	类型	技术运用	特点
Chinese skill	综合	语音合成技术	多针对初级学习者，有详细的拼音发音教程和初级词汇，主题偏生活化
Hello Chinese	综合	语音识别技术	过关类学习 APP，每一关卡设置听、说、读、写
Super Chinese	综合	深度学习技术	可根据自身情况制定学习计划，同时根据用户水平调整内容难度

中文学习APP	类型	技术运用	特点
Pleco	词典	文字识别技术	页面简洁、翻译准确，可自己创建词卡
Pinyin News	商务、阅读	语音识别技术	实时更新、简明概述新闻内容
Skritter	汉字；写作	文字识别技术	提供汉字练习的同时兼顾语音、语义方面的练习
Art of Chinese	汉字	手写识别技术	中国风页面，设计感较强，提供汉字的演化过程
正音万里行	拼音	语音识别技术	专门针对声调和拼音学习,有发音讲解，内容完整
Hello Daily	综合	语音识别技术	分课学习
嗨中文	口语	语音识别技术	“短视频+直播”
M-Madarin	综合	手写识别技术	漫画学汉语，对话场景丰富
SPK Chinese	综合	深度学习技术	私人订制学习

目前中文学习 APP 各具特点，从呈现方式来看，中文学习 APP 有文本、图片、音频、视频、动画、注释、故事、游戏、对话等形式，呈现方式总体上较为丰富。从页面设计来看各具特色，很多 APP 设计中融入了中文传统文化元素，如熊猫、书法等。从功能上来看，大部分 APP 设置了练习测试，学习者针对某一主题或专项学习后，对学习内容进行检测。部分 APP 中设置了一定的奖励机制，激励学习者提高软件学习使用率，如“Chinese skill”、“Super Chinese”等。一些 APP 注重用户的情感体验功能，如通过社交互动提高用户体验。但是现有 APP 中只有少部分设计了互动功能，交互功能的呈现以批改作业为主，提供问答互动的 APP 数量较少。另外，只有少量 APP 设计

了评测功能。

从技术应用方面来看，当前语音识别技术在中文学习 APP 中应用广泛，如 Hello Chinese、正音万里行、Hello daily、嗨中文、e 学中文、Chinese skill 等 APP 都可以实现学习者录音、评测功能。语音合成技术是人机交互的关键，在 APP 中的应用如 Chinese skill 等。文字识别技术在中文学习 APP 中的应用也逐渐成熟，Pleco、Skritter、Art of Chinese 等汉字学习 APP 基本都已具备文字识别功能，且准确率较高。深度学习技术在 APP 中的应用尚不广泛，Super Chinese 和 SPK Chinese 采用了深度学习技术，根据大数据和学习者自身情况，实现个性化的学习方案的制定。

以 SPK Chinese 为例，其通过 AI 人工智能技术，实现私人定制学习，包含语音、图文识别以及汉语翻译。适用于来华汉语学习者、参加 HSK 考试的学生、来中国旅游、商务合作的学习者等。其特色如下：

（1）SPK Chinese 根据学习者兴趣推荐中文阅读文章，每日为其推荐中国新闻、谚语故事、武侠小说等阅读内容。

（2）具备 AR 查词功能，学习者可以手机拍照扫描文本，语音识别翻译，也可以通过语音识别功能纠正学习者的发音。

（3）提供 HSK 考试 AI 学习中心，基于 HSK1-6 级考试题库，通过覆盖中国本地生活场景的汉语对话进行模拟培训。

（4）社群合作学习，全球学习者可以实现随时随地交友聊天。

（5）结合大数据挖掘和分析实现私人定制，学习者可以每日查看学习进度报告。

9.5.2 中文教学平台

在众多的国际中文教学平台中，具有较大规模影响力的有全球中文学习平台[438]、中文联盟（网络孔子学院）[439]、唐风汉语国际教育云平台[440]、长城汉语智慧云平台[441]、国际中文智慧教育平台、Ponddy

Reader（庞帝智能中文教学平台）[442]等。此外，近年来许多国内教育科技类公司纷纷布局国际中文教育产业，创建了包括哈兔中文网络学院[443]、锦灵中文[444]、悟空中文[445]、Lingo Ace[446]、Lingo Bus[447]、PPtutor[448]、Chinlingo[449]等一系列网络在线中文教学平台。这些教学平台的基本信息见表 9-2。

当前，国际中文教学平台在研发和构建时普遍都遵循着整体性原则、灵活性原则、个性化原则和资源集成原则。整体性原则指的是国际中文教学平台普遍覆盖了“课前、课中、课后”完整的教学环节并拥有集“教、学、测、评、管”等于一体的功能。灵活性原则指的是国际中文教学平台在设计时充分考虑了教师和学生主体之间的需求差异。个性化原则是指国际中文教学平台可以根据教师和学习者的不同需求，为其提供一些列精细化网络教学工具或针对性地为学习者推送学习知识和练习题目，尽可能地提高教学效率和学习效率。资源集成原则指的是国际中文教学平台除了实现“教”与“学”的功能外，还致力于为教师和学生两大教学主体提供丰富多样的教学资源和学习资源，最大可能地满足教师的教学需求和学生的学习需求。

表 9-2：部分国际中文教学平台基本信息介绍

国际中文教学平台	创建时间	创建研发单位	主要面向对象	平台主要特点
全球中文学习平台	2019 年	科大讯飞股份有限公司	全球中文学习者 国内各大相关高校	运用智能语音技术、人工智能技术，可为学习者制定个性化学习方案，提供多样的学习资源和应用，建有在线教学平台，遵循“共建共享”的原则
中文联盟	2020 年	五洲汉风网络科技（北京）有限公司等	全球中文学习者 中外国际中文教育机构	建有在线教学平台，可提供“教、学、考、研”一体化服务，并配套了齐全的教学、课程、教材、测试等资源

国际中文教学平台	创建时间	创建研发单位	主要面向对象	平台主要特点
唐风汉语国际教育云平台	2006年	北京唐风汉语教育科技有限公司研发	国内外相关高校和机构	利用视频结构化技术、大数据技术，集教学教务、教学资源、测评、教育应用于一体，可实现个性化教学
长城汉语智慧云平台	2005年	北京汉雅天诚教育科技有限公司研发	国内外相关高校和机构	采用人工智能技术、网络多媒体技术，课前、课中、课后闭环式教学，终端、平台、应用、数据深度融合，教、学、管、测、评全面衔接
国际中文智慧教育平台	2022年	北京语言大学	国内外相关高校和机构	利用大数据技术、人工智能技术，以融课件和智能习题本为核心，可形成用户学习自画像，强调移动学习、互动学习、个性化学习和自适应学习
Ponddy Reader	2018年	庞帝智能科技（北京）有限公司创建	中文学习者、中文教师 海内外相关机构	运用人工智能技术、自然语言处理技术，开发了多个智能技术工具，拥有丰富的教学、学习、教材等资源，可为教师备课、教学和学习者自主学习提供便捷
哈兔中文网络学院	2012年	杭州哈兔网络科技有限公司创建	海外华侨华人子女	可实现个性化、定制化教学，自主研发了一套课程体系，倡导互动式、趣味性教学
锦灵中文	2018年	北京赛酷雅科技有限公司创建	青少儿	研发了独特的课程体系，注重文化教学，善于利用动画、电子绘本、音频等多媒体资源
悟空中文	2015年	上海凯晏教育科技有限公司创建	海外青少儿（3-18岁）	自主研发了分级教材和课程，倡导探究式学习法和“全身反应法”理念，注重全面培养汉语各项技能，重视文化教学
Lingo Ace	2017年	武汉领格教育科技有限公司创建	青少儿（3-15岁）	注重互动和沉浸式教学，课程贴合相关大纲标准，课堂拥有丰富的教学活动，强调闭环学习

国际中文教学平台	创建时间	创建研发单位	主要面向对象	平台主要特点
Lingo Bus	2017年	北京谦育科技有限公司创建	全球青少儿（4-15岁）	依据相关大纲和标准设计课程，倡导重视沉浸式教学和文化教学，运用了支架式教学等多种教学理念
PPtutor	2017年	成都学语教育科技有限公司创建	华裔青少儿（4-15岁）	研发了不同的课程体系，倡导教学的互动性、趣味性和完整闭环，实现了个性化教学
Chinlingo	2014年	厦门中学西渐信息科技有限公司创建	全球汉语和中国文化爱好者	有着丰富的课程资源，并与相关大纲标准对标，教学形式多样化，可实现个性化教学

9.6 总结和展望

目前，智能技术正向汉语教学各相关领域内部渗透，教学和研究与技术的融合日趋加深。未来智能技术将从以下三个方面持续对国际中文教育产生巨大的影响。

1. 新基建：智能技术赋能国际中文教育数字基础设施建设

智能技术正在深刻而广泛地改变着国际中文教育。在教学资源方面，智能技术改变了国际中文教学资源的面貌，丰富了教学资源的类型与模态；在教学实践方面，智能技术加强了课堂教学的互动性和学生学习的自主性；在语料库方面，人工智能、云计算、计算机自动标注等先进技术先后被运用到语料库建设中，生成技术的应用为超大规模语料库建设提供了可能；在综合应用方面，应用多种智能技术手段的智慧教学平台不断发展，逐渐涵盖课前、课中、课后各个环节，为管理者、教师、学生等多种身份的参与者提供更便捷的服务。

未来，国际中文教育必须做好顶层设计，做好国际中文教育相关数据和过程的标准化工作，以智能技术为驱动，加强建设国际中文教育数字基础设施，打破国际中文教育在全球发展不平衡的局面，借助

智能技术赋能在不同国家和区域实现国际中文教育资源共享。

2. 新业态：智能技术催化汉语国际教育产学研结合

随着智能技术在国际中文教育领域的不断深入，无论是基础设施建设还是工程应用实践都对软硬件、人才、资金等各方面提出了越来越高的要求。基于此，我们认为国际中文在线教育建设应该继续秉承“共建共享”的原则，倡导广大高校、科研单位和社会力量齐发力，共同参与研发实践；加快培养一批具有国际中文教育视野和掌握智能技术的复合型人才，为国际中文在线教育提供智力支持；统筹发展国际中文教育事业和国际中文教育产业，加强产学研互动，打造具有创新性和实用性的汉语国际教育产品，将语言教育与文化、技术、经济进行深度的融合。

3. 新模式：智能技术推动汉语国际教育数字化转型

在未来的一段时间内，如何将国际中文教育与“互联网+”深度融合仍是一项重要的课题。进入“十四五”以来，国家相继出台了一系列建设“数字中国”、数字经济、数字社会的规划，旨在加快信息化、数字化与国民经济的深度融合。在此背景下，中外语言合作交流中心于 2021 年 12 月发布了《国际中文在线教育行动计划（2021-2025 年》[450]，“计划”指出了发展国际中文在线教育的重要意义和现实必要性，并从标准与机制的构建、相关平台建设、相关资源和课程资源建设等 6 个方面提出了远景规划，到 2025 年要基本实现国际中文教育数字化、智能化和泛在化的发展目标。

本章编写人员：

王治敏、王一帆、赵慧周、汪张龙、杨冰冰、袁亮杰、徐悦

第 10 章 多语种智能信息处理团队

10.1 民族语言信息处理及语料库建设

10.1.1 内蒙古大学

内蒙古自治区蒙古文信息处理技术重点实验室于 2007 年经内蒙古自治区科技厅批准在内蒙古大学挂牌成立，实验室主任为飞龙教授。重点实验室长期专注于蒙古文智能信息处理研究和蒙古文软件及平台开发工作，通过高水平的科研工作培养蒙古文信息处理领域的高层次人才，为少数民族语言文字信息化的发展提供了人才保障和技术支持，推动了蒙古文信息处理技术的发展。实验室先后主持承担了国家 863 和 973 计划专项课题，在国内外代表性期刊和学术会议发表学术论文 200 多篇，其中在 TALSP、ICASSP 等国际著名学术期刊和会议发表论文 100 余篇。实验室研发了蒙古语语音识别与合成、新蒙古文与传统蒙古文相互转换、蒙汉翻译等一系列智能系统与软件。

内蒙古大学蒙古语文研究所是内蒙古大学蒙古学研究领域中最早建立的研究机构之一。1982 年，内蒙古大学蒙古语文研究所成立。1995 年，内蒙古大学蒙古学研究院（后改为蒙古学学院）成立，蒙古语文研究所成为隶属该研究院的一个单位。该研究所研究人员围绕现代蒙古语及方言、中世纪蒙古语及蒙古语文献、北方民族古文字、蒙古语族及阿尔泰语系、蒙古文信息处理、实验语音学、社会语言学与文化语言学等研究方向开展研究工作。

10.1.2 西北民族大学

西北民族大学甘肃省民族语言智能处理重点实验室聚焦国家重大需求，围绕语言智能、文化计算、社会与环境计算、计算核医学展开研究。“语言智能”方向围绕多语言信息技术、社会发展所面临的重大科学问题，秉承以多语言信息为主导的新经济产业化理念，进一步拓展多语言信息技术研究的领域，进行成果转化和社会服务，更好

地维护民族团结和边疆稳定和“一带一路”建设。“文化计算”方向响应国家战略需求，助力中华文化遗产保护。“社会与环境计算”方向以大数据技术为基础，开展社会计算和环境计算两个方面的工作，服务于国家社会发展和生态文明建设。“计算核医学”方向响应核能创新国家战略需求，助力人民生命安全和健康。

西北民族大学民族信息技术科研团队开创性地研制藏文系列软件。在激烈地国际竞争中，国际标准化组织通过了以中国提案为主的藏文编码国际标准，藏文成为我国第一个具有国际编码的少数民族文字，从科学技术领域证明了中国政府为捍卫人权而付出的努力。团队研制了世界上第一个藏文视窗平台、字处理软件和藏文网站，载入了中国人权白皮书。在以汉字为核心的多语言人工智能技术领域不断取得一系列重大科技成果。西北民族大学民族信息技术科研团队荣获中共中央组织部、中国中央宣传部、人力资源和社会保障部、科学技术部颁发的“全国专业技术人才先进集体”殊荣。

西北民族大学中国民族语言文字信息技术教育部重点实验室是西北民族大学在诸多领域开展关于藏文的多语言智能处理研究建成的重点实验室。重点实验室围绕以国家通用语为核心的多语言智能处理、脑科学、知识可视化、文化遗产数字化保护等领域，面向国家战略需求、“一带一路”建设开展工作。重点实验室围绕以国家通用语为核心的多语言智能处理、脑科学、知识可视化、文化遗产数字化保护等领域，面向国家战略需求、“一带一路”建设开展工作。重点实验室主要研制藏文操作系统，各种通用应用软件，从藏文编码国际标准、字型国家标准，到 WINDOS 环境下多语言机器翻译、语音识别、语音合成、文字识别、舆情分析、知识图谱、社交媒体、情感分析、语言认知等。

10.1.3 新疆大学

新疆多语种信息技术重点实验室依托于新疆大学，其前身是新疆

大学多语种信息技术重点实验室，2007 年经自治区人民政府批准成立“新疆民文信息技术研发中心”，2008 年 12 月由新疆维吾尔自治区科学技术厅批准成立“新疆多语种信息技术重点实验室”。重点实验室主任是吾守尔·斯拉木院士。重点实验室以多语种信息处理及系统软件为总的研究发展方向；以国家丝绸之路经济带建设、网络空间安全、自治区信息化为向社会服务方向，发挥在信息处理技术和中西亚多语言学科交叉的优势和特色。

新疆大学新疆多语种信息技术研究中心作为国家语委的科研基地，是国家语委组织高水平科学研究、汇聚和培养优秀科研人才、加强学科建设、开展学术交流、提供语言服务的重要平台，是提升语言文字工作决策科学化水平、加强语言文字管理工作的重要保障。研究中心主任为吾守尔·斯拉木院士，副主任为杨文忠教授。研究中心主要开展新疆少数民族语言和中西亚国家多种语言多模态资源和标准建设、多种自然语言处理、语音识别、语言合成、语音翻译、语音人机交互等技术研究及语言文字信息技术的社会服务工作。

10.1.4 中科院新疆理化技术研究所

中国科学院新疆理化技术研究所多语种信息技术研究室主要从事多语种信息处理关键技术研究、电子政务和电子商务关键技术及平台研究、多语种软件的测试、研究。近年来，已完成 2 项国家“863”和 1 项科技攻关项目，在多语种信息处理技术与电子政务方面开发出维哈柯文永中 office 办公套件，维哈柯文信息发布系统，西北星信息发布系统和西北星电子政务应用基础平台。

10.1.5 中央民族大学

“国家语言资源监测与研究少数民族语言中心”由国家教育部语信司、国家民委教科司、国家新闻出版总署报刊司及中央民族大学共同创建。机构设立在中央民族大学中国少数民族语言文字信息化工程研究中心，主任是赵小兵教授。中心自 2008 年成立以来承担的研究

课题和研究内容与少数民族语言监测研究及应用密切相关，包括蒙、藏、维、哈、彝等民族语言信息处理相关的基础语料库、知识库构建，编码转换、分词技术等基础信息处理技术研究，以及跨语言社会舆情分析研究等。已承担国家科技支撑计划项目 1 项、国家自然科学基金重点课题等 4 项、国家社科基金项目 1 项、国家新闻出版“十一五”重大科研项目 1 项、国家语委等省部级重点项目 4 项、国家语委和国家民委等省部级一般项目 3 项；获国家发明专利 3 项、软件著作权 11 项；发表学术论文 99 篇，出版学术专著或教材 13 部，开发并完成成果转化的软件系统 13 项等。

10.2 东盟语言信息处理及语料库建设

10.2.1 阿里巴巴达摩院

阿里巴巴达摩院语言技术实验室是阿里巴巴负责 NLP 技术研发的核心团队，为阿里巴巴经济体提供包括 NLP 基础技术、对话技术、应用算法、机器翻译、内容搜索推荐等技术。达摩院语言技术实验室负责人是司罗，实验室在多语言技术方面上聚焦多语言和跨语言技术领域，如东南亚语基础 NLP、跨领域学习、自监督学习、低资源 NLP 等。实验室在多语言 NER、泰语越南语分词、情感分析/地址解析等多语言技术上相对比较成熟。截止 2021 年，语言技术实验室在各大国际顶级学术会议发表论文 160 余篇，获权威竞赛奖项 20 余个，并创造了多项技术突破及应用创新。

10.2.2 广东外语外贸大学

广东外语外贸大学现有 28 个外语语种，是华南地区外语语种最多和高层次外语人才最多的高校。2019 年，广州市非通用语种智能处理重点实验室获得广州市科学技术局批准并依托广东外语外贸大学建设运行，其主任是蒋盛益教授。实验室面向“一带一路”和“粤港澳大湾区”战略的重大需求，充分利用和发挥广东外语外贸大学在华南地区的非通用语种研究与人才培养优势，聚焦“一带一路”互联

互通、语言互通的多语种智能信息处理，兼顾当前需求与长远发展，以突破人工智能应用基础理论瓶颈为重点，重点围绕非通用语种信息处理技术、基础资源建设、非通用语言教学以及服务国别区域研究为目标导向，研究数据驱动与知识引导的非通用语种人工智能新方法，旨在打造粤港澳大湾区非通用语种智能信息处理研究的标杆团队，助力广州与“一带一路”沿线国家的文化交流和经贸往来，同时为广东企业走出去和“粤港澳大湾区”建设提供非通用种语言服务和技术支撑。

10.2.3 昆明理工大学

昆明理工大学云南省海量语言信息处理工程实验室充分发挥云南省数字经济开发区南亚东南亚语言教育独有资源和多语种人才优势，帮助园区打造多语种软件研发、多语系互译应用、多文化优势输出的特色工程。通过“语言技术”的应用和产业合作平台建设，研发出“云岭翻译”及“小语洞听”、“小语洞见”、“小语会议”、“小语译制”等“小语智能”成果。工程实验室的主任是余正涛教授，实验室的主要研究方向为：面向南亚东南亚小语种自然语言处理、多语言机器翻译、跨语言信息检索及舆情分析。

10.2.4 南宁市平方软件新技术有限责任公司

南宁市平方软件新技术有限责任公司（平方软件）负责人是刘连芳教授，主要围绕东南亚语言处理、少数民族语言处理及多媒体等技术进行研究，并开发出电子政务软件、少数民族文化软件、多语言辅助翻译软件等产品。

10.3 多语种语法分析、翻译及语料库建设

10.3.1 爱丁堡大学

爱丁堡大学自然语言处理小组（EdinburghNLP）致力于研究使计算机能够理解和产生人类语言的算法。该小组的研究涉及自然语言处理的所有核心领域，包括形态学、句法分析、语义、语篇、语言生成

和机器翻译。EdinburghNLP 在 NLP 的接口与其他方面也有很多的研究，包括语音技术、机器学习、计算机视觉、认知模型、社会媒体、信息检索、机器人、生物信息学、教育技术。小组的成就包括神经机器翻译系统 Nematus 和高性能语言建模工具包 KenLM。

10.3.2 澳门大学

澳门大学自然语言处理与中葡机器翻译实验室致力于研究葡萄牙语和汉语之间不同的语系和拓扑结构，并进一步对葡萄牙语和汉语之间的机器翻译进行研究。截止 2020 年，自然语言处理与中葡机器翻译实验室在国际期刊、国内期刊及较有影响的国际会议上发表论文 100 余篇，举办了五场中葡机器辅助系统和原型机的新闻发布会。

10.3.3 巴斯克大学

巴斯克大学 IXA 的工作范围从计算语言学的基础研究到人类语言技术的关键应用，涵盖信息检索和信息提取、机器翻译、语言学习、形态学、句法-形态语法、词典学语义、基本设施和语言工具（SGML、XML）的集成等领域。他们的研究为巴斯克语提供了强大的，宽覆盖的自然语言处理的技术。这些技术包括拼写检查器，机器翻译系统，Basque Wordnet，科学技术语料库和语法注释的语料库。

10.3.4 百度研究院

百度研究院，隶属于百度公司，归属于百度 AI 技术平台体系，下设认知计算实验室、硅谷人工智能实验室、深度学习实验室、大数据实验室、商业智能实验室、量子计算研究所、机器人与自动驾驶实验室、安全实验室和生物计算实验室。2013 年初组建了深度学习研究院，即百度研究院的前身。2014 年，百度研究院正式成立。2017 年 3 月，百度成立 AI 技术平台体系(AlG)，由王海峰博士担任总负责人。百度研究院的研究方向包括机器学习、数据挖掘、计算机视觉、语音、自然语言处理、商业智能、量子计算等。吴华博士担任百度自然语言处理首席科学家，研究院自然语言处理领域涉及的研究内容包

括：NLP 基础算法、机器学习基础技术、语义计算、语言理解、语言生成、问答系统、对话系统和机器翻译。截止 2021 年，百度自然语言和语音领域在各大国际顶级学术会议发表论文 100 余篇。

10.3.5 北京大学

北京大学计算语言研究所从事的研究任务包括计算语言学、自然语言处理、也包括机器学习、深度学习、人工智能等相关前沿领域研究。目前计算语言研究所挂靠在北京大学信息科学技术学院，现任所长为王厚峰教授，副所长为穗志方教授及詹卫东教授。研究所围绕计算语言学和自然语言处理，研究包括如下三个主要的方向：基础理论、NLP 的模型和方法包括计算语言学基础，自然语言处理核心技术，现代汉语语法，汉语的词/句法/语义分析，NLP 统计模型，语言处理的信息论方法等；基础资源的研究与建设包括计算词典学与机器词典，综合型语言知识库，语料库语言学与语料库加工技术，术语学、术语自动提取、术语标准化研究等；基础应用技术包括机器翻译的方法、技术与系统实现，信息检索与提取，自然语言信息处理系统的评价方法和技术，受限汉语及其辅助写作系统，中国古诗词计算机辅助研究等。重点研究课题涉及的内容有：语言模型与分析技术，深度学习技术，信息检索与提取的模型与系统，计算语义学，自然语言处理系统评价技术，自动文本摘要，机器翻译的理论、技术与系统实现，问答系统，机器阅读理解。

10.3.6 北京交通大学

北京交通大学语言智能与大数据处理研究所负责人是徐金安教授，研究所在自然语言处理、机器翻译、知识图谱及其应用、文本情感分析、自动摘要、问答、对话系统、人机交互等方面都有丰硕的研究成果。尤其是在机器翻译研发方面，多次在国内外的机器翻译评测任务中夺得第一名，其中，WAT2016 夺得 3 个第一名、CCMT2020 夺得 4 个第一名、WMT2021 夺得英汉翻译的第一名、CCMT2021 夺

得 5 个第一名。

10.3.7 北京理工大学

北京理工大学北京市海量语言信息处理与云计算应用工程技术研究中心于 2009 年批准设立，主任为黄河燕教授。工程中心主要研究方向包括机器翻译、海量数字资源管理、机器学习与 Web 挖掘、信息检索与社会计算、语义计算与知识工程、云计算应用与安全等。工程中心自认定以来，成功申请并承担了国家 973 计划课题、国家 863 计划课题、国家科技支撑计划项目、国家自然科学基金重点项目、国家重点研发计划项目等国家重大/重点项目 30 余项，并荣获国家科技进步二等奖、北京市科学技术一等奖、三等奖等多项省部级及以上科研奖励。

北京理工大学语言智能与社会计算团队主要研究方向是机器翻译和自然语言处理。团队学术带头人是黄河燕教授，黄教授目前还兼任教育部计算机专业教指委副主任委员、信息技术新工科产学研联盟副理事长兼秘书长、北京市海量语言信息与云计算应用工程技术研究中心主任。团队成员多次在人工智能与自然语言处理领域顶级国际期刊和学术会议发表学术论文，并有部分博士学位论文被中国中文信息学会或中国人工智能学会评为优秀博士学位论文。

北京理工大学 NLPIR 大数据搜索与挖掘实验室的负责人张华平副教授，实验室面向海量异构新型社交媒体互联网，研究网络大数据搜索、多语言自然语言处理、知识图谱与社会舆情分析等关键技术，服务于国家安全治理、行业大数据挖掘与个人智能服务。实验室目前承担了国家自然科学基金、973 计划、863 计划、242 课题、两高一部等国家课题 20 余项，新疆自治区高新技术计划、河北省科技支撑计划等省部级课题 3 项。实验室核心成果 NLPIR 多语大数据语义增强分析平台覆盖了中文、英文、西班牙语、法语、维语、阿拉伯语、印度乌尔都语、多哥语等“一带一路”沿线语言的自然语言处理，搭

建了 NLPIR 大数据语义增强分析平台，融合了 NLPIR-ICTCLAS 汉语分词系统、JZSearch 大数据搜索引擎、KGB 知识图谱引擎、九眼智能过滤等工具。

10.3.8 重庆大学

重庆大学语料库研究所以语料库为平台进行语言理论及应用研究，现任所长为李良炎副教授。语料库研究所的研究特色包括：语料库语言学理论与技术研究（针对语料库建设中句法标注与语义标注难题，依托人工智能、机器学习、数据挖掘等计算机前沿技术进行攻关）；专门用途英语语类研究及语料库建设（主要研究基于系统功能语法理论建立专门用途英语语类的理论框架，并对专门用途英语语类问题，尤其是学术英语语料库建设以及辅助学术写作软件开发进行深入研究）；自然语言处理技术研究（提出了基于词联接的自然语言处理技术，并在中国古典诗词语言处理系统中得到验证）。

10.3.9 传神语联网网络科技股份有限公司

传神语联网网络科技股份有限公司（Transn 传神）负责人是何恩培，主要围绕人工智能、大数据、互联网等技术开展基于场景的多语信息服务的研究，重点围绕产能组织调度技术、人机共译技术和机器翻译技术进行研发，并开创了“人机共译”的新型语言产能“第三产能”—Twinslator。

10.3.10 德国人工智能研究中心

德国人工智能研究中心（DFKI）研究方向覆盖人工智能的主要产业方向，包括大数据分析、知识管理、画面处理和理解和自然语言处理、人机交互、机器人。研究中心负责人是 Wolfgang Wahlster 教授。该研究中心分别在柏林的语音和语言技术实验室以及在位于萨尔布吕肯的多语言与语言技术部进行语言方面的研究。其研究的中心领域是机器学习、文本分析、机器翻译、自然语言对话系统、人机交互以及数字内容的创建和管理。

10.3.11 东北大学

东北大学自然语言处理实验室在姚天顺教授和王宝库教授带领下于 1973 年创立，先后由朱靖波教授、肖桐教授领导。东北大学自然语言处理实验室长期从事自然语言处理领域的研究，涉及机器翻译、语言分析、文本处理等多个方向，在自然语言处理和人工智能领域重要国际会议和知名期刊上发表论文 200 余篇。近年来，实验室在人工智能领域顶级会议及期刊上（如 ACL、AI）上发表学术论文 70 余篇，研发了 NiuTrans、NiuTensor、NiuParser 等多个开源系统，并曾在 WMT、CCMT/CWMT、NTCIR PatentMT 等国内外机器翻译评测任务中获得冠军。其中 NiuTrans 开源机器翻译系统形成了支持 304 种语言互译的机器翻译应用平台。

10.3.12 东京工业大学

东京工业大学 Okazaki 实验室主要研究人工智能中的自然语言处理，负责人是 Naoaki Okazaki 教授。实验室目前正在探索在计算机上实现智能交流的原则和方法，如翻译外语文本、与人交流、回答问题和解释场景。截止 2022 年，实验室在各大国际顶级学术会议发表论文 80 余篇。此外，实验室还开发所研究内容的实际应用，例如，使用大数据分析的社交倾听。

10.3.13 Facebook 人工智能研究院

Facebook 人工智能研究院研究的系统可以对日常交流中常见的非正式语气、俚语和拼写错误保持不错的识别能力。该团队的研究人员致力于深度学习/神经网络、自然语言处理、语言识别、文本规范化、词义消歧和机器学习等复杂问题，以分解问题，并构建和部署健壮的语言翻译解决方案。

10.3.14 哥伦比亚大学

哥伦比亚大学自然语言处理研究小组负责人是 Michael Collins 教授，小组研究的领域包括：音韵学与韵律，语法和分析，词汇语义，

词义消歧义，话语处理，话语共指，对话与口语，方言变化，信息提取、数据挖掘，机器翻译，语言和社交网络，阿拉伯语自然语言处理。研究小组在全球顶尖的会议期刊上发表论文有 590 余篇，创建了语料库、词典等多种语言资源。

10.3.15 Google Research

Google Research 的自然语言处理研究侧重于大规模、跨语言和跨域应用的算法，机器翻译的研究侧重于开发统计翻译技术。机器翻译使用了大规模的计算基础设施，可以对基于网络规模数据训练的新模型进行快速试验，从而提高翻译质量。截止 2022 年 6 月，Google Research 的自然语言处理团队和机器翻译团队在各大国际会议及期刊上共发表论文 900 余篇。

10.3.16 哈尔滨工业大学

哈尔滨工业大学计算机学院机器智能与翻译研究室一直致力于机器翻译研究与系统开发，学术带头人为博士生导师李生教授和赵铁军教授。实验室研究方向包括：互联网信息智能处理；机器翻译及多语言信息处理；故事理解与动画生成技术；语言分析技术与应用和机器学习；人工智能技术的应用。研究室完成我国第一个通过技术鉴定的汉英机器翻译系统 CEMT-I，获部级科技进步二等奖。2000 年以来先后承担国家自然科学基金课题 8 项；863 项目 7 项；承担国防项目 2 项，国家信息安全中心项目 3 项，国际合作项目 6 项以及多项省部级科研课题和横向课题。开发英汉双语树库 2 万多句对、中英日三语语料库 7 万句对等，并在汉语分词及词性标注、词义消歧、中英文句法分析、信息检索、多文档文摘、话题检测与跟踪、文景转换、语音系统、知识工程、测井曲线处理等诸多方面进行了有特色的研究，取得了一批有一定影响的成果。近五年来，实验室在国际期刊、国内重要期刊及较有影响的国际会议上发表论文 200 余篇。

哈尔滨工业大学社会计算与信息检索研究中心（HIT－SCIR）成

立于 2000 年 9 月，研究中心主任是刘挺教授。研究中心主要研究方向包括句子级的语言分析（句法分析、语义分析、命令解析、文本顺滑）、人机对话（营销/客服机器人、深度问答、用户画像与机器人画像、话题推荐）、篇章级的自然语言理解与生成（阅读理解、篇章语义、信息抽取、语言知识图谱、文本生成、新闻自动写作）、社会计算（倾向性分析、观点分析、情绪分析、消费意图识别、事理图谱、社会预测）。研究中心积极参加国内外技术评测并取得优异成绩，包括国际 CoNLL 2009 七国语言句法语义分析评测、CoNLL 2018 国际多语言通用依存分析评测第一名、CoNLL 2019 国际跨框架语义分析评测第一名。已完成或正在承担的国家 973 课题、国家自然科学基金重点项目、国家 863 重点项目、国际合作等课题 60 余项。近年来持续在自然语言处理和人工智能国际顶级会议上发表多篇高水平论文。

哈尔滨工业大学智能技术与自然语言处理研究室由王晓龙教授担任负责人，是国内较早从事自然语言处理研究的科研团体之一。自八十年代初期以来，先后开展了俄汉机器翻译、固定段落问答、自动文摘、文本纠错、汉字智能输入、语音识别与合成、语料库多级加工、语言模型、信息检索、问答系统等多项研究。研究室的代表性成果是开创性地提出了汉字语句输入的思想并实现了国内外第一个语句级汉字键盘输入系统。目前共获得部科技进步级一等奖 1 项，二等奖 4 项，获得国家专利 3 项。先后在国内外重要学术刊物和会议上发表论文 200 余篇，编著书 8 部。1990 年以来完成的国家自然科学基金重点/面上项目、国家 863 重点/面上项目、中美、中日国际合作等重要科研项目 20 多项。

10.3.17 哈佛大学

哈佛大学 NLP 研究小组致力于研究处理和生成人类语言的机器学习方法，小组负责人是 Stuart M. Shieber 教授。小组对序列生成的数学模型、基于人类语言的人工智能的挑战以及使用统计工具探索语

言结构等领域感兴趣。小组的研究集中于文本摘要、神经机器翻译、可视化递归神经网络、收缩神经网络算法、文档实体跟踪模型、多模式文本生成、语法错误纠正和文本生成新方法。

10.3.18 合肥工业大学

合肥工业大学情感计算研究所于2011年成立，主任为孙晓教授。研究所主要从事先进智能、情感计算、大规模数据与知识获取的基础理论研究工作。研究所构建了丰富的数据资源，包括文本语料库、面部表情库、动作-情感库等，其中全球最大规模的中文情感语料库，已授权全球近300家高校、科研机构使用。构建了基于大规模多源情感数据库的情感感知、推理与交互的总体研究思路与方法体系，取得了系列突破性研究成果，使我国在先进智能及情感机器人达到了世界先进水平。主持并参与国家973预研项目、863项目，国家自然科学基金项目、安徽省自然科学基金项目、企业委托项目等多项。2011年12月获批建设“情感计算与先进智能机器安徽省重点实验室”，系统展开情感计算与先进智能机器的研究，是国内首个以情感计算命名的重点实验室。研究方向包括：情感计算；自然语言处理；听觉信息认知计算；视觉信息认知计算；情感可穿戴计算；机器人云理论及其应用。

10.3.19 华南师范大学

华南师范大学自然语言处理与智能软件技术研究团队主要从事自然语言处理和人工智能软件技术的研究、设计、开发与应用，团队负责人是曾碧卿教授。研究团队的代表性成果是在文本情感分析研究中设计了局部与全局上下文特征提取器，显著提升了方面级情感分析的效果。研究团队已经发表相关科研学术及其教育教改论文100余篇，出版学术专著2部，主编出版专业教材13部，申请发明专利12项。研究团队目前的主要研究方向：文本情感分析、推荐系统、聊天机器人、智能软件机器人、自动文本摘要、机器阅读理解、机器翻译、知

识图谱、知识推理、问答系统、强化学习、多轮对话、问题生成、实体关系抽取，以及自然语言处理在教育中应用的关键技术研究等。

10.3.20 华为诺亚方舟实验室

实验室成立于2012年，实验室的研究领域主要包括计算机视觉、自然语言处理、搜索与推荐、决策与推理、人工智能理论等。其中语音和语言处理方面由刘群教授担任首席科学家。语音和语言处理方面研究课题包括：大规模预训练语言模型及应用，NLP模型压缩和加速、语音识别与合成、机器翻译、对话系统、自然语言生成和问答，AI同声传译等。截止2020年，华为“诺亚方舟实验室”语音和语言处理方面在各大国际顶级学术会议发表论文60余篇。

10.3.21 剑桥大学

剑桥大学自然语言和信息处理研究小组致力于计算语言学，自然语言处理和信息检索。该小组的研究项目包括语言处理资源和工具、逻辑和形式、前端以及语音处理。最近的项目涉及工具和处理器的进一步开发、自动摘要、文本和语音信息检索；形式规范的自然语言处理；词汇知识的获取和多语词汇知识库的构建。

10.3.22 卡耐基梅隆大学

卡耐基梅隆大学语言技术研究所（LTI）由Jaime Carbonell教授创建，研究所在自然语言处理、计算语言学、信息抽取、摘要和问题回答、信息检索、文本挖掘和分析、知识表示、推理和获取、教育语言技术、机器学习、机器翻译、多模式计算和交互、语音处理、口语接口和对话处理等领域进行开创性的研究。

10.3.23 科大讯飞认知智能国家重点实验室

2017年12月13日，科技部批准依托科大讯飞股份有限公司建设认知智能国家重点实验室。实验室重点开展语义计算、知识建模等认知智能基础理论及技术研究。实验室自成立以来，在国际权威的评测和竞赛中获得二十余项世界冠军，承担国家及省部级科技项目近二

十项。此外，重点实验室研发还研发出人机语音交互、多语种语音翻译、智能客服、个性化学习、智医助理等认知智能系统。

10.3.24 马里兰大学

马里兰大学 CLIP 实验室致力于设计算法和方法，实验室主任是 Jordan Boyd-Graber 教授。CLIP 实验室研究涵盖了语言计算研究的主要领域，包括但不限于深度学习、多语言文本处理、机器翻译、计算心理语言学和语音检索以及跨语言信息检索。

10.3.25 慕尼黑大学

慕尼黑大学信息和语言处理中心对自然语言处理及其理论基础进行跨学科研究，中心负责人是 Hinrich Schütze 教授。中心使用的主要方法是以语言学为基础的统计 NLP，研究的问题包括计算语法和语义、情感分析、机器翻译和半监督学习、词汇资源的改编和扩展。该组织已经在相关的核心期刊和学术会议上发表论文近 200 篇，并为大多数欧洲语言以及中文和韩语创建了最大的德语电子词典以及词典。

10.3.26 南加州大学

南加州大学信息科学研究所(ISI)是一家隶属于安德鲁和埃尔纳维特比工程学院学术研究机构。南加州大学自然语言处理小组是南加州大学计算语言学社区的一部分，小组在自然语言处理和计算语言学等方面进行了大量研究。小组的主要研究方向为：机器翻译、机器问答、文本摘要、信息检索以及自然语言生成。

10.3.27 南京大学

南京大学自然语言处理研究组曾先后承担过该领域的 18 项国家科技攻关项目、863 项目、国家自然科学基金和江苏省自然科学基金以及多项对外合作项目的研制。其中，包括承担的国家七五科技攻关项目“日汉机译系统研究”。近年来在陈家骏教授带领下，南京大学自然语言处理研究组集中关注文本分析、机器翻译、社交媒体分析推

荐、知识问答等多个热点问题，结合统计方法和深度学习方法进行问题建模和求解，取得了丰富的成果。在自然语言处理顶级国际会议和人工智能顶级国际会议上发表论文三十余篇，相关系统在机器翻译、中文分词、命名实体识别、情感计算等多个国际国内评测中名列前茅。

10.3.28 欧洲语言资源协会

欧洲语言资源协会（ELRA）主要任务是使人类语言技术的语言资源向广大社区开放。协会主要研究领域：跨文化交际、欧洲语言、信息和通信技术、语言障碍、语言技术、多种语言、战略指导。

10.3.29 清华大学

清华大学自然语言处理与社会人文计算实验室在20世纪70年代末由黄昌宁教授的带领下成立，实验室主要从事自然语言处理的研究工作，是国内开展相关研究最早、深具影响力的科研单位，是中国中文信息学会（全国一级学会）计算语言学专业委员会挂靠单位。主持国家项目及基金16项，其中包括国家863计划两项和国家973计划一项。自然语言处理与社会人文计算实验室主要在自然语言处理方向上进行研究，其中机器翻译和知识计算方面尤为突出。团队研发自动作诗系统九歌，支持 11 种语言的 THUMT 开源机器翻译工具包及OpenNRE神经网络关系抽取工具包等。

10.3.30 斯坦福大学

斯坦福大学自然语言处理小组的工作范围从计算语言学的基础研究到人类语言技术的关键应用，小组负责人是Christopher Manning教授。小组研究内容涵盖句子理解，自动问答，机器翻译，句法解析和标记，情感分析，对话代理，文本和视觉场景的情绪，以及自然语言处理在数字人文和计算社会科学中的应用等领域。团队开发出Stanford CoreNLP（集成NLP工具包），词性标记器，命名实体识别器，以及处理60多种人类语言文本的Stanza工具包。

10.3.31 苏州大学

苏州大学自然语言处理实验室由周国栋教授组建，以自然语言理解、中文信息处理、机器翻译和自然语言认知为主要研究方向。实验室围绕自然语言处理领域的发展趋势，强调原创性科学研究与应用开发研究相结合，已成为我国自然语言处理领域的重要研究基地和具有国际影响力的研究中心之一。近 5 年在自然语言处理和人工智能的国际顶级 SCI 期刊和 CCF A/B 类会议上发表论文 120 多篇，获得国家自然科学基金杰青项目 1 项、重点项目 3 项、面上青年项目 30 多项，国家重点研发计划课题 2 项、子课题 2 项，国际重大合作项目 1 项。

10.3.32 台湾大学

台湾大学自然语言处理实验室主要的研究方向是人类语言技术、资讯检索与撷取、网路探勘、和人工智慧，实验室负责人是陳信希教授。实验室目前正研究的课题包括：结合微观与巨观之跨语言跨文件知识发掘；生活纪录；知识图谱；网路意见探勘；使用者意图与行动之分析和预测；机器翻译；自动摘要和问答；相关性、多样性和新颖性资讯之分析、侦测和追踪；社会媒体检索；情境导向资讯检索；跨语言跨媒体资讯检索；学习式排序法在资讯检索上的应用。

10.3.33 腾讯人工智能实验室

腾讯人工智能实验室（AI Lab）的在自然语言处理方面研究重点是增强自然语言中计算机与人之间的相互作用，研究涵盖文本理解、文本生成、对话和机器翻译。史树明博士担任腾讯 AI Lab 自然语言处理中心负责人。截止 2021 年，AI Lab 自然语言处理方面在各大国际顶级学术会议发表论文 240 余篇。除此之外自然语言处理方面还开发出智能创作助手 Effidit，文本理解系统 TexSmart 以及交互翻译系统 TranSmart。

10.3.34 天津大学

天津大学智能与计算学部自然语言处理团队负责人是熊德意教

授，团队专注于机器翻译、对话、自然语言生成、问答与机器阅读理解、信息抽取与知识图谱、认知启发的 NLP 等方向的研究，在国际著名期刊和会议上发表论文 100 余篇，Springer 出版英文专著一部，编著会议论文集多部。获得国家自然科学基金优秀青年科学基金（国家优青）、国家重点研发计划“政府间国际科技合作创新合作”重点专项、英国皇家学会牛顿高级学者基金、国家留学基金委“双一流”大学建设高校专项创新型人才国际合作培养等项目资助。

10.3.35 拓尔思信息技术股份有限公司

拓尔思信息技术股份有限公司总裁是施水才教授，公司主要围绕大数据、人工智能、互联网内容管理、网络信息安全和互联网营销等领域进行研究，开发出了海贝大数据管理系统，智拓语义智能技术平台，基于深度学习的自然语言处理引擎以及智能问答机器人等产品。

10.3.36 微软亚洲研究院

1998 年 7 月，李开复加入微软并在中国创建并领导微软中国研究院。微软亚洲研究院自然语言计算组专注于自然语言处理领域的理论、模型、算法和应用的研究和创新。目前主要的研究领域包括：自然语言理解与生成，机器翻译，智能问答，语音处理，代码智能，文档智能，多模态理解与生成，以及大规模预训练模型等。

10.3.37 厦门大学

厦门大学智能科学与技术系自然语言处理实验室的负责人是史晓东教授。目前团队的三个主要研究方向是多语种机器翻译、隐喻计算、信息抽取和检索。团队面向信息社会中不同语言间的国内外交流的重大需求，以实现高度智能化的语言处理为目标，开展分词、命名实体识别、句法分析、多语词语对齐、语言模型、隐喻理解等基础研究，规则和统计相结合的机器翻译、辅助翻译、嵌入式翻译、新型语篇语义翻译模型、云翻译平台、神经机器翻译等机器翻译研究，知识图谱、信息抽取、关系和事件抽取、跨语言信息检索、舆情分析等网

络信息处理应用基础研究。主要成果包括古汉字数据库，“云译”平台，神经机器翻译系统，简繁文本智能转换平台，“云使”跨语言搜索引擎，异体字词典。

10.3.38 香港科技大学

香港科技大学人类语言科技中心（HLTC）是香港科技大学的一个多学科研究中心，中心负责人是吴德恺教授。该中心由来自电子工程师学会和计算机科学系的教职员工领导，专门从事语音和信号处理、统计和基于语料库的自然语言处理、机器翻译、文本挖掘、信息抽取、中文处理、知识管理和相关领域。HLTC 建立的系统包括互联网自动语言翻译、基于语音的网络浏览和电话语音识别。

10.3.39 西湖大学

西湖大学文本智能实验室是西湖大学工学院以张岳课题组为基础设立的。西湖大学文本智能实验室致力于研究自然语言处理技术的基础问题，以算法研究为主，同时涉猎认知、脑科学、神经科学、量子计算等方向，探索能够主动学习、可解释、稳定理解和生成人类语言的计算模型。研究包括基础的词法、句法、语义理解，信息抽取中的命名实体、关系、事件和情感分析，文本生成中的数据评论、文本摘要、机器翻译、问答对话系统等任务，同时不断探索最先进的语言处理技术对于跨学科的医疗、金融等领域的帮助作用。

10.3.40 约翰霍普金斯大学

约翰霍普金斯大学语言与语音处理中心专注于语言和语音的科学和技术，其负责人是 Andreas Andreou 教授。该中心的研究涉及广泛的基础和应用课题，包括声学处理、自动语音识别、大数据、认知建模、计算语言学、信息抽取、机器学习、机器翻译和文本分析。

10.3.41 郑州大学

郑州大学自然语言处理实验室成立于 2004 年 10 月，隶属于郑州大学信息工程学院。在咎红英教授的带领下，实验室从成立至今在现

代汉语广义虚词库建设、文本自动分类、中文文本的褒贬评价、中英文双语术语抽取等方向进行了深入研究，并在该领域积累了丰富的阶段性成果。实验室目前在人工智能领域重要期刊和会议上发表学术论文 150 余篇，参与或承担国家级项目 9 项，省部级项目 26 项。

10.3.42 中译语通科技股份有限公司

中译语通科技股份有限公司（中译语通）的首席技术官是程国艮，公司主要进行语言科技与服务、大数据与人工智能的研究，自主研发了机器翻译、跨语言大数据分析、金融量化与监管科技、全球科技发现与价值评估、数字城市大脑和工业互联网等系统平台。拥有覆盖机器翻译、自然语言处理、跨语言大数据分析和知识图谱等领域的自主知识产权技术体系。

10.1.43 字节跳动人工智能实验室

字节跳动人工智能实验室成立于 2016 年，其中自然语言处理邻域的研究课题包括：句法和语义分析、情感分析、文本分类、文本匹配和检索、文本摘要、对话系统、问答、机器翻译、自然语言生成、信息提取以及语言和视觉。自然语言处理邻域的负责人是李航博士。截止 2021 年，自然语言处理研究课题组开发出的应用有 Byte Translator，AI 写稿机器人 Xiaomingbot 以及头条和抖音的搜索服务。

10.4 国际中文教育智能处理

10.4.1 北京语言大学

北京语言大学汉语国际教育研究院是北京语言大学在教育部人文社会科学重点研究基地“对外汉语研究中心”与汉语国际推广基地“国际汉语教学研究基地”基础上，重新整合校内外汉语国际教育学科优势资源与优秀人才，全新打造的一个旗舰级汉语国际教育研究机构。院长是吴应辉教授，副院长是王治敏教授和姜丽萍教授。研究院将学术研究与智库研究并重，在十二个重点领域开展研究：一是面向汉语国际教育的语言本体研究；二是语言学习与认知研究；三是互动

语言学与汉语教学研究；四是汉语国际教育资源研发；五是汉语国际教育技术应用研究；六是汉语教学理论与方法相关研究；七是汉语国际教育发展与评价研究；八是孔子学院及世界主要语言传播机构发展研究；九是汉语国际传播的区域与国别研究；十是汉语国际教育数据库建设与研究；十一是汉语国际教育语料库建设与研究；十二是汉语国际教育信息资源集成与服务。

北京语言大学语言信息处理研究所始建于 1987 年，是中国境内第一个以汉语信息处理为主要研究方向的研究所。研究所的人员由语言学、计算机软硬件、信息处理等领域的专家和教师组成。第一任所长是马希文教授，第二任所长是张普教授，第三任所长是宋柔教授。该所的宗旨是：发展汉语信息处理技术，用以支持语言本体研究和对外汉语教学，以及该领域的其他应用工作。

本章编写人员：

孙晓、徐豪杰、刘维锋、刘江维、史云伟、赵旭阳

第 11 章 参考文献

[1] 伊.达瓦, 句坂芳典, 卢绪刚,等. 蒙古语连续语音识别在不同结构语言模型下精度的讨论[C]// 全国人机语音通讯学术会议. 2009.

[2] 伊·达瓦, 大川茂树, 白井克彦. 蒙古语多方言语音识别及共享识别模型探索[J]. 中央民族大学学报(哲学社会科学版), 2001(4):114-121.

[3] 包世恩. 蒙古语非特定人大词汇量连续语音识别系统的研究与实现[D]. 呼和浩特: 内蒙古大学, 2005.

[4] Gao G L, Zhang S. A Mongolian speech recognition system based on HMM[C]//Proc of International Conference on Intelligent Computing. 2006: 667-676.

[5] Qilao H, Gao G L. Researching of speech recognition oriented mongolian acoustic model[C]//Proc of 2th Pattern Recognition, 2008. CCPR'08. Chinese Conference. 2008: 1-6.

[6] Bao F, Gao G L. Improving of acoustic model for the mongolian speech recognition system[C]//Proc of 2th Pattern Recognition CCPR, 2009: 1-5.

[7] 飞龙，高光来，王宏伟.基于词干的蒙古语语音关键词检测方法的研究[J].中文信息学报，2016，30(1) ： 124-128.

[8] Bao F, Gao G L, Yan X, et al. Segmentation-based Mongo-lian LVCSR approach[C]//Proc of 38th ICASSP, 2013: 1-5.

[9] Zhang H, Bao F, Gao G L. Mongolian speech recognition based on deep neural networks[C]//Proc of 15th Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, 2015: 180-188.

[10] Zhang H W, Bao F, Gao G L, et al. Comparison on Neural Network based acoustic model in Mongolian speech

recogni-tion[C]//Proc of 20th Asian Language Processing (IALP), 2016 International Conference, 2016: 1-5.

[11] Y. Wang, F. Bao, H. Zhang and G. Gao, "Joint Alignment Learning-Attention Based Model for Grapheme-to-Phoneme Conversion," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7788-7792, doi: 10.1109/ICASSP39728.2021.9413679.

[12] T. Zhi, Y. Shi, W. Du, G. Li and D. Wang, "M2ASR-MONGO: A Free Mongolian Speech Database and Accompanied Baselines," 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 2021, pp. 140-145, doi: 10.1109/O-COCOSDA202152914.2021.9660401.

[13] Y. Qian and Z. Zhou, "Optimizing Data Usage for Low-Resource Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 394-403, 2022, doi: 10.1109/TASLP.2022.3140552.

[14] 李振宏，高光来. 印刷体蒙古文文字识别中常用特征的获取[J]. 计算机技术与发展，2003，11): 117-9.

[15] 李振宏，高光来，侯宏旭，et al. 印刷体蒙古文文字识别的研究 [J]. 内蒙古大学学报:自然科学版，2003: 454-7.

[16] 魏宏喜. 印刷体蒙古文字识别中关键技术的研究[D]; 内蒙古大学，2006.

[17] 魏宏喜，高光来. 印刷体蒙古文字识别中蒙古文字特征的选择[J]. 内蒙古大学学报:自然科学版，2006，37(006): 694-7.

[18] Gao G，Su X，Wei H，et al. Classical Mongolian words recognition in historical document[C]//2011 International Conference

on Document Analysis and Recognition. IEEE，2011: 692-697.

[19] 苏向东. 蒙古文古籍识别技术的研究[D]. 内蒙古大学，2011.

[20] Su X，Gao G，Wang W，et al. Character segmentation for classical Mongolian words in historical documents[C]//Chinese Conference on Pattern Recognition. Springer，Berlin，Heidelberg，2014: 464-473.

[21] Su X，Gao G，Wei H，et al. Enhancing the Mongolian historical document recognition system with multiple knowledge-based strategies[C]//International Conference on Neural Information Processing. Springer，Cham，2015: 536-544.

[22] Su X，Gao G，Wei H，et al. A knowledge-based recognition system for historical Mongolian documents[J]. International Journal on Document Analysis and Recognition (IJDAR)，2016，19(3): 221-235.

[23] Daoerji F，Guanglai G. DNN-HMM for large vocabulary Mongolian offline handwriting recognition[C]//2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE，2016: 72-77.

[24] Zhang H，Wei H，Bao F，et al. Segmentation-free printed traditional Mongolian OCR using sequence to sequence with attention model[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE，2017，1: 585-590.

[25] Wang W，Wei H，Zhang H. End-to-end model based on bidirectional lstm and ctc for segmentation-free traditional Mongolian recognition[C]//2019 Chinese Control Conference (CCC). IEEE，2019: 8723-8727.

[26] Wei H，Gao G. A holistic recognition approach for woodblock-print Mongolian words based on convolutional neural network[C]//2019 IEEE International Conference on Image

Processing (ICIP). IEEE，2019: 2726-2730.

[27] Kang Y，Wei H，Zhang H，et al. Woodblock-printing Mongolian words recognition by bi-LSTM with attention mechanism[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE，2019: 910-915.

[28] Wei H，Liu C，Zhang H，et al. End-to-end model for offline handwritten Mongolian word recognition[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer，Cham，2019: 220-230.

[29] 刘聪. 大词汇量脱机手写蒙古文整词识别研究[D].呼和浩特：内蒙古大学，2019.

[30] 李敏. 基于深度学习的联机蒙古文手写识别系统研究[D].呼和浩特：内蒙古大学，2019.

[31] Fan D，Gao G，Wu H. Sub-word based Mongolian offline handwriting recognition[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE，2019: 246-253.

[32] 范道尔吉. 蒙古文脱机手写识别研究[D]. 内蒙古大学，2020.

[33] Wei H，Zhang H，Zhang J，et al. Multi-task learning based traditional Mongolian words recognition[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE，2021: 1275-1281.

[34] Cui S D，Su Y L，Ji Y T. An end-to-end network for irregular printed Mongolian recognition[J]. International Journal on Document Analysis and Recognition (IJDAR)，2022，25(1): 41-50.

[35] Wei H，Liu K，Zhang J，et al. Data Augmentation Based on CycleGAN for Improving Woodblock-Printing Mongolian Words Recognition[C]//International Conference on Document Analysis and Recognition. Springer，Cham，2021: 526-537.

[36] Zhang H，Chen W，Su X，et al. An Efficient Local Word Augment Approach for Mongolian Handwritten Script Recognition[C]// International Conference on Document Analysis and Recognition. Springer，Cham， 021: 429-443.

[37] Su X，Xu H，Zhang Y，et al. An End-to-End Preprocessor Based on Adversiarial Learning for Mongolian Historical Document OCR[C]//Pacific Rim International Conference on Artificial Intelligence. Springer，Cham，2019: 266-272.

[38] Su X，Xu H，Kang Y，et al. Improving text image resolution using a deep generative adversarial network for optical character recognition[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE，2019: 1193-1199.

[39] 敖其尔. 一种波形拼接的语音合成实验[A]. 见：第三届全国人机语音通讯学术会议（NCMMSC1994）论文集[C]. 1994. 408-412.

[40] 高光来，孟和，吉雅. 基于词汇的蒙古语文语转换的实验[J]. 内蒙古大学学报: 自然科学版, 2000, 31(1): 121-124.

[41] 萨其容贵. 蒙古语语音合成技术的研究[D]. 呼和浩特: 内蒙古大学，2005.

[42] 田会利. 基于词干词缀的有限词条的蒙古语语音合成系统的研究[D]. 呼和浩特: 内蒙古大学, 2007.

[43] 孟和吉雅, 敖其尔. 基于词干词缀的蒙古语语音合成方法[J]. 内蒙古大学学报: 自然科学版, 2008, 39(6): 693-697.

[44] 敖敏. 基于韵律的蒙古语语音合成研究[D]. 内蒙古大学, 2012.

[45] 李婷会. 蒙古语的韵律预测方法研究[D]. 内蒙古大学, 2014.

[46] 刘瑞. 基于条件随机场的蒙古语韵律短语预测方法[A]. 见：中国中文信息学会语音信息专业委员会.第十三届全国人机语音通讯学术会议(NCMMSC2015)论文集[C].中国中文信息学会语音信

息专业委员会:清华信息科学与技术国家实验室(筹),2015. 552-556.

[47] Liu R, Bao F, Gao G, et al. Mongolian prosodic phrase prediction using suffix segmentation[C]. In: Proceedings of the 2016 International Conference on Asian Language Processing (IALP). IEEE, 2016. 250-253.

[48] 赵建东, 高光来, 飞龙. 基于 HMM 的蒙古语语音合成技术研究[J]. 计算机科学, 41(1): 80-82.

[49] 赵建东. 基于隐马尔科夫模型的蒙古语语音合成技术研究[D]. 内蒙古大学, 2014.

[50] Liu R, Bao F, Gao G, et al. Mongolian text-to-speech system based on deep neural network[C]. In: Proceedings of the 2017 National Conference on Man-Machine Speech Communication (NCMMSC). Springer, Singapore, 2017. 99-108.

[51] Li J, Zhang H, Liu R, et al. End-to-End Mongolian Text-to-Speech System[C]. In: Proceedings of the 2018 International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2018. 483-487.

[52] 刘郅楠. 基于端到端蒙古语语音合成方法的研究[D]. 内蒙古大学, 2019.

[53] Liu R, Bao F, Gao G. Building Mongolian TTS Front-End with Encoder-Decoder Model by Using Bridge Method and Multi-view Features[C]. Proceedings of the 26th International Conference on Neural Information Processing (ICONIP2019), 2019: 642-651.

[54] Liu R, Sisman B, Li J, et al. Teacher-Student Training for Robust Tacotron-Based TTS[C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

IEEE, 2020: 6274-6278.

[55] 刘瑞. 基于深度学习的蒙古语语音合成研究[D]. 内蒙古大学, 2020.

[56] Liu R, Sisman B, Bao F, et al. Exploiting Morphological and Phonological Features to Improve Prosodic Phrasing for Mongolian Speech Synthesis [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29(99):274-285

[57] Liu R, Sisman B, Bao F, et al. Modeling prosodic phrasing with multi-task learning in tacotron-based TTS[J]. IEEE Signal Processing Letters, 2020, 27: 1470-1474.

[58] Liu R, Sisman B, Gao G, et al. Expressive TTS Training with Frame and Style Reconstruction Loss [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1806-1818.

[59] Liu R, Sisman B, lai Gao G, et al. Decoding Knowledge Transfer for Neural Text-to-Speech Training[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30:1789-1802.

[60] 武子玉. 基于半监督方法的蒙汉机器翻译的研究[D]. 内蒙古大学, 2020.

[61] 苏依拉, 贺玉玺, 王昊,等. 一种基于数据增强的蒙汉神经机器翻译方法[P].中国，202011580153，2020.

[62] 吉亚图. 低资源神经机器翻译中关键问题的研究[D]. 内蒙古大学, 2020.

[63] Nier Wu, Hongxu Hou, Ziyue Guo, Wei Zheng. Low-Resource Neural Machine Translation Using XLNet Pre-training Model[C]// 30th International Conference on Artificial Neural Networks (Volume 12895). 2021:503-514.

[64] 白天罡. 基于强化学习的蒙汉神经网络机器翻译的研究[D].

内蒙古大学, 2020.

[65] 白慧琨,王斯日古楞,宁静.基于条件随机场的蒙古文人名识别[J].内蒙古师范大学学报(自然科学汉文版),2016,45(02):253-255.

[66] 哈斯高娃,王斯日古楞.基于条件随机场模型的蒙古文地名自动识别研究[J].内蒙古师范大学学报(自然科学汉文版),2019,48(01):82-85.

[67] 才晶晶.基于CRF的蒙古文人名自动识别[D].内蒙古大学,2016.

[68] 吴金星,那顺乌日图,杨振新.基于 CRF 的蒙古文人名自动识别研究[J].计算机应用研究,2016,33(07):2014-2017.

[69] 包乌格德勒,鲍薇.基于条件随机场的蒙古文地名识别[J].现代计算机(专业版),2017(03):6-9.

[70] 吴金星,丽丽,杨振新.CRF 和词典相结合的蒙古文地名识别研究[J].计算机工程与科学,2016,38(05):1046-1051.

[71] 王玉荣,林民,李艳玲.BERT 蒙古文词向量学习[J/OL].计算机工程与应用:1-7, 2022.

[72] 王炜华.蒙古文命名实体识别研究[D].内蒙古大学,2018.

[73] 吴都.基于深度神经网络的蒙古文命名实体识别研究[D].北京交通大学,2020.

[74] 熊玉竹.融合语言模型和注意力机制的蒙古文命名实体识别研究[D].内蒙古大学,2019.

[75] 安苏雅拉,基于 transformer 神经网络的汉蒙机构名翻译研究,中文信息学报，第 34 卷第 1 期，2020 年 1 月.

[76] 王玉荣，林民，李艳玲，BERT 跨语言词向量学习研究，计算机科学与探索，2021-04.

[77] 王辉.“一带一路”国家语言状况与语言政策.北京：社会科学文献出版社，2015.33-168.

[78] 王晋军,施黎辉.中国与东盟国家民族语言政策对比研究.北京：

社会科学文献出版社，2020.49-188

[79] Pisceldo F, Mahendra R, Manurung R, Arka I W. A Two-Level Morphological Analyser for the Indonesian Language[C]// Proceedings of the Australasian Language Technology Association Workshop 2008. 2008: 142-150.

[80] Larasati S D, Kuboň V, Zeman D. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus[C]// International Workshop on Systems and Frameworks for Computational Morphology. Springer, Berlin, Heidelberg, 2011: 119-129.

[81] Sodhy G C. Prefix Extraction of Malay Words using Backpropagation Neural Network [R]. UTMK technical report, UTMK, USM, 1998.

[82] Sulaiman S, Gasser M, Kübler S. Towards a Malay Derivational Lexicon: Learning Affixes Using Expectation Maximization[C] //Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011. 2011: 30–34.

[83] Gunawan W, Suhartono D, Purnomo F, Ongko A. Named-Entity Recognition for Indonesian Language using Bidirectional LSTM-CNNs[J]. Procedia Computer Science, 2018, 135: 425-432.

[84] Wibawa A S, Purwarianti A. Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning[J]. Procedia Computer Science, 2016, 81:221-228.

[85] Alfred R, Leong L C, On C K, Anthony P. Malay named entity recognition based on rule-based approach[J]. International Journal of Machine Learning & Computing, 2014, 4(3):300-306.

[86] Irmawati B, Shindo H, Matsumoto Y. A Dependency Annotation

Scheme for Indonesian[C]//The 21st Annual Meerting of The Association for Natural Language Processing for Japan. Japan, 2015: 740–743.

[87] Gusmita R H, Manurung R. Some initial experiments with Indonesian probabilistic parsing[C]//Proceedings of the 2nd International MALINDO (Malay and Indonesian Language) Workshop. 2008.

[88] Joice. Pengembangan lanjut pengurai struktur kalimat bahasa indonesia yang menggunakan constraint-based formalism. Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2002. Call number: SK-0487.

[89] Irmawati B, Shindo H, Matsumoto Y. A dependency annotation scheme to extract syntactic features in Indonesian sentences[J]. International Journal of Technology, 2017, 8(5): 957-967.

[90] Herlim R S, Purwarianti A. Indonesian Shift-Reduce Constituency Parser Using Feature Templates & Beam Search Strategy[C]//2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA). IEEE, 2018: 54-59.

[91] Rahman A, Purwarianti A. Ensemble Technique Utilization for Indonesian Dependency Parser[C]// Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. 2017: 64-71.

[92] Abidin A I Z, Yong S P, Kasbon R, Azman H. Utilizing Top-down Parsing Technique in the Development of a Malay language Sentence Parser[C] //Second International Conference on Informatics. 2007: 125–131.

[93] Noor Y M, Jamaludin Z. Parser with Sentence Correction for

Malay Language (BM)[J]. 2012 International Conference on Information and Knowledge Management (ICIKM 2012), 2012, 45: 138–142.

[94] Hiloh M A F, Ab Aziz M J, Zakaria L Q. The Effectiveness of Bottom Up Technique with Probabilistic Approach for A Malay Parser[J]. GEMA Online® Journal of Language Studies, 2018, 18(2): 124-133.

[95] Noor N H M, Sapuan S, Bond F. Creating the Open Wordnet Bahasa[C]// Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation. 2011: 255–264.

[96] Mahendra R, Septiantri H, Wibowo H A, Manurung R, Adriani M. Cross-Lingual and Supervised Learning Approach for Indonesian Word Sense Disambiguation Task[C]//Proceedings of the 9th Global WordNet Conference (GWC 2018). 2018: 248-253.

[97] Final Report on Statistical Machine Translation for Bahasa Indonesia-English and English-Bahasa Indonesia[R]Agency for the Assessment and Application of Technology (Badan Pengkajian dan Penerapan Teknologi, BPPT). 2009.

[98] Hermanto A, Adji T B, Setiawan N A. Recurrent neural network language model for English-Indonesian Machine Translation: Experimental study[C]//Proceedings of the 2015 International Conference on Science in Information Technology. IEEE, 2015: 132–136.

[99] Yeong Y L, Tan T P, Gan K H, Mohammad S K. Hybrid Machine Translation with Multi-Source Encoder-Decoder Long Short-Term Memory in English-Malay Translation[J]. International Journal on Advanced Science, Engineering and Information Technology, 2018,

8(4-2): 1446-1452.

[100] Wang P D, Nakov P, Ng H T. Source Language Adaptation Approaches for Resource-Poor Machine Translation[J]. Computational Linguistics, 2016, 42(2): 277–306.

[101] Octoviani W, Fachrurrozi M, Yusliani N, Febriady M, Firdaus A. English–Indonesian Phrase Translation Using Recurrent Neural Network and ADJ Technique[C]//Journal of Physics: Conference Series. IOP Publishing, 2019, 1196(1): 012007.

[102] Yusoff N, Jamaludin Z, Yusoff M H. Semantic-based Malay-English translation using n-gram model[J]. Journal of Telecommunication, Electronic and Computer Engineering, 2016, 8(10): 117–123.

[103] Yeong Y L, Tan T P, Mohammad S K. Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System[J]. Procedia Computer Science. 2016, 81: 243–249.

[104] 郑锫涛，林楠铠，付颖雯，王连喜，蒋盛益.汉语-印尼语平行语料自动对齐方法研究. [J]. 《广西师范大学学报》(自然科学版), 2019, 37(1): 89-97.

[105] Qiu X Y, Zhu G Q. Learning Indonesian-Chinese Lexicon with Bilingual Word Embedding Models and Monolingual Signals[C]// Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016). 2016: 188-193.

[106] Soleh M Y, Purwarianti A. A non word error spell checker for Indonesian using morphologically analyzer and HMM[C]// Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, IEEE, 2011: 1-6.

[107] Irmawati B, Shindo H, Matsumoto Y. Exploiting Syntactic Similarities for Preposition Error Corrections on Indonesian Sentences Written by Second Language Learner[J]. Procedia Computer Science. 2016, 81: 214–220.

[108] Fahda A, Purwarianti A. A Statistical and Rule-Based Spelling and Grammar Checker for Indonesian Text[C]//2017 International Conference on Data and Software Engineering. IEEE, 2017: 1-6.

[109] Mawardi V C, Susanto N, Naga D S. Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method[C]//MATEC Web of Conferences. EDP Sciences, 2018, 164: 01047.

[110] Lin N K, Chen B Y, Wattanachote K, Jiang S Y. A Framework for Indonesian Grammar Error Correction. 2019.

[111] Kasbon R, Amran N A, Mazlan E M, Mahamad S. Malay Language Sentence Checker[J]. World Applies Sciences Journal 12, 2011, 12: 19–25.

[112] Basri S B, Alfred R, On C K. Automatic Spell Checker for Malay Blog[C]//2012 IEEE International Conference on Control System, Computing and Engineering. IEEE, 2012: 506–510.

[113] Noor Y M, Jamaludin Z. PARSE TREE VISUALIZATION FOR MALAY SENTENCE (BMTutor)[J]. ARPN Journal of Engineering and Applied Sciences, 2015, 10(3): 1253–1259.

[114] Franky, Bojar O, Veselovská K. Resources for Indonesian Sentiment Analysis[J]. The Prague Bulletin of Mathematical Linguistics, 2015, 103(1): 21–41.

[115] Koto F, Rahmaningtyas G Y. InSet Lexicon: Evaluation of a Word List for Indonesian Sentiment Analysis in Microblogs[C]//

Proceedings of the 21st International Conference on Asian Language Processing (IALP 2017). 2017: 391–394.

[116] Lunando E, Purwarianti A. Indonesian social media sentiment analysis with sarcasm detection[C]//2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2013). IEEE, 2013: 195–198.

[117] Fauzi M A. Random Forest Approach fo Sentiment Analysis in Indonesian[J]. Indonesian Journal of Electrical Engineering and Computer Science, 2018, 12(1): 46-50.

[118] Sadanandan A A, Osman N A, Saifuddin H, Ahamad M K, Pham D N, Hoe H. Improving Accuracy in Sentiment Analysis for Malay Language[C]//4th International Conference on Artificial Intelligence and Computer Science. 2016: 28–29.

[119] Al-Saffar A, Awang S, Tao H, Omar N, Al-Saiagh W, Al-bared M. Malay sentiment analysis based on combined classification approaches and Senti-lexicon algorithm[J]. PloS one, 2018, 13(4): e019485.

[120] Koto F. A Publicly Available Indonesian Corpora for Automatic Abstractive and Extractive Chat Summarization[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2016: 801-805.

[121] Kurniawan K, Louvan S. IndoSum: A New Benchmark Dataset for Indonesian Text Summarization[C]//Proceedings of the International Conference on Asian Language Processing (IALP 2018). 2018:215-220.

[122] Cai Z F, Lin N K, Ma C Y, Jiang S Y. Indonesian Automatic Text Summarization based on A New Clustering Method in Sentence Level

[C]//Proceedings of the 2019 International Conference on Big Data Engineering. 2019: 30-35.

[123] Roxas R, Mula G. A morphological analyzer for Filipino verbs[C]//Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation. 2008: 467-473.

[124] Bonus D E. The Tagalog Stemming Algorithms (TagSA)[C]//the Proceedings of the Natural Language Processing Research Symposium, DLSU, Manila. 2003.

[125] Erlyn M, Yuji M. Factors Affecting Part-of-Speech Tagging for Tagalog[C]//Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2. 2009: 763-770.

[126] Go M P, Nocon N. Using Stanford Part-of-Speech Tagger for the Morphologically-rich Filipino Language[C]//Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. 2017: 81-88.

[127] Clark A. Unsupervised induction of stochastic context-free grammars using distributional clustering[C]//Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7. Association for Computational Linguistics, 2001: 13.

[128] Alcantara D L, Borra A. Constituent Structure for Filipino: Induction through Probabilistic Approaches[C] //Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation. 2008: 113-122.

[129] Manguilimotan E, Matsumoto Y. Dependency-based Analysis for Tagalog Sentences[C]//Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation. 2011: 343-352.

[130] Mistica M, Baldwin T. Recognising the predicate-argument structure of Tagalog[C] //Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. 2009: 257-260.

[131] Borra A. A Transfer-Based Engine for an English to Filipino Machine Translation Software[D]. MS Thesis, University of the Philippines Los Baños, 1999.

[132] Roxas R. A Hybrid English-Filipino Machine Translation System[C]//3rd National Natural Language Processing Research Symposium. DLSU-Manila. 2006.

[133] Roxas R E O, Borra A, Cheng C K, et al. Building language resources for a Multi-Engine English-Filipino machine translation system[J]. Language resources and evaluation, 2008, 42(2): 183-195.

[134] Ong E, Go K, Nuñez V A, et al. Template-based English-Filipino machine translation system[C]//Proceedings of the 4th National Natural Language Processing Research Symposium. 2007.

[135] Ang J, Chan M R, Genato J P, et al. Development of a Filipino-to-English Bidirectional Statistical Machine Translation System that dynamically updates via user feedback[C]//Proceedings of the 12th International Workshop on Spoken Language Translation, Da Nang, Vietnam. 2015.

[136] Tacorda A J, Ignacio M J, Oco N, et al. Controlling byte pair encoding for neural machine translation[C]//2017 International Conference on Asian Language Processing (IALP). IEEE, 2017: 168-171.

[137] Lazaro A N，Oco N，Roxas R E. Developing a Bidirectional

Ilocano-English Translator for the Travel Domain：Using Domain Adaptation Techniques on Religious Parallel Corpora[C]//11th International Conference of the Asian Association for Lexicography. Guangzhou：Guangdong University of Foreign Studies，2017：889.

[138] Andrei A L. Development and Evaluation of Tagalog Linguistic Inquiry and Word Count (LIWC) Dictionaries for Negative and Positive Emotion [EB/OL]. [2019-12-30]. https://www.mitre.org/sites/default/files/publications/pr_14-3858-development-evaluation-of-tagaloglinguistic-inquiry.pdf.

[139] Regalado R V J, Chua J L, Co J L, et al. Subjectivity Classification of Filipino Text with Features Based on Term Frequency--Inverse Document Frequency[C]//2013 International Conference on Asian Language Processing. IEEE, 2013: 113-116.

[140] PIPPIN M，ODASCO R，DE JESUS R，et al. Classifications of Emotion Expressed by Filipinos Through Tweets[C]// Proceedings of the International Multi Conference of Engineers and Computer Scientists. Hong Kong：[s. n.]，2015：18-20.

[141] Lapitan F R, Batista-Navarro R T, Albacea E. Crowdsourcing-based annotation of emotions in filipino and english tweets[C]// Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016). 2016: 74-82.

[142] Eboña K M L, Llorca Jr O S, Perez G P, et al. Named-entity recognizer (NER) for Filipino novel excerpts using maximum entropy approach[J]. Journal of Industrial and Intelligent Information Vol, 2013, 1(1).

[143] Alfonso A P T，Domingo I V R，Galope M J F，et al. Named Entity Recognizer for Filipino Text Using Conditional Random Field[J].

International Journal of Future Computer and Communication，2013，2(5)：376.

[144] Dimalen E D, Dimalen D M D. An open Office spelling and grammar checker add-in using an open Source External Engine as Resource Manager and Parser[C]//Proceedings of the 4th National Natural Language Processing Research Symposium (NNLPRS), CSB Hotel. 2007.

[145] Oco N, Borra A. A grammar checker for Tagalog using LanguageTool[C]//Proceedings of the 9th Workshop on Asian Language Resources. 2011: 2-9.

[146] Go M P，Nocon N，Borra A. Gramatika：A Grammar Checker for the Low-Resourced Filipino Language[C]//TENCON 2017-2017 IEEE Region 10 Conference. Penang：IEEE，2017：471-475.

[147] Tsao N L，Wible D. A Method for Unsupervised Broad-Coverage Lexical Error Detection and Correction[C]//Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications. Morristown：Association for Computational Linguistics，2009：51-54.

[148] Huang C C，Chen M H，Huang S T，et al. EdIt： A Broad-Coverage Grammar Checker Using Pattern Grammar[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies：Systems Demonstrations. Morristown ： Association for Computational Linguistics，2011：26-31.

[149] Tiu E P, Roxas R E. Automatic Bilingual Lexicon Extraction for a Minority Target Language[C]//Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation. 2008:

368-376.

[150] Dita S, Roxas R E, Inventado P. Building online corpora of philippine languages[C]//Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2. 2009: 646-653.

[151] Dita S，Roxas R E，Inventado P. Building Online Corpora of Philippine Languages[C]//Proceedings of the 23rd Pacific Asia Conference on Language，Information and Computation. Hong Kong：City University of Hong Kong，2009：646-653.

[152] Chu S. Language Resource Development at DLSUNLP Lab[C/OL]//The School of Asian Applied Natural Language Processing for Linguistics Diversity and Language Resource Development ADD-4：Language Resource Technology. Bangkok：[s.，2009.[2019-12-30]. http://xsite.dlsu.edu.ph/research/centers/adric/nlp/downloads/(ADD4)%20Shirley-Language%20Resource%20Development%20at%20DLSU-NLP%20Lab.pdf.

[153] Borra A, Pease A, Roxas R, et al. Introducing Filipino WordNet[C]//Principles, Construction and Application of Multilingual Wordnets: Proceedings of the 5th Global WordNet Conference. 2010.

[154] El-Kishky A, Chaudhary V, Guzman F, et al. A Massive Collection of Cross-Lingual Web-Document Pairs[J]. arXiv preprint arXiv: 1911.06154, 2019.

[155] Sagum R A, Ramos A D, Llanes M T. FICOBU: Filipino WordNet Construction Using Decision Tree and Language Modeling[J]. International Journal of Machine Learning and Computing, 2019, 9(1).

[156] Kexiao Zheng, Wenkui Zheng, "Deep Neural Networks Algorithm for Vietnamese Word Segmentation", Scientific Programming, vol. 2022, Article ID 8187680, 11 pages, 2022.

[157] Chen, Z. et al. (2018). Vietnamese Part of Speech Tagging Based on Multi-category Words Disambiguation Model. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Y. (eds) Natural Language Processing and Chinese Computing. NLPCC 2017. Lecture Notes in Computer Science, vol 10619. Springer, Cham. https://doi.org/10.1007/978-3-319-73618-1_23

[158] Vu T, Nguyen D Q, Dai Q N, et al. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2018.

[159] Quach L D, Thanh D D, Tran D C, et al. Comparative Study of Vietnamese Part-of-Speech Tagging Tools[C]// Comparative Study of Vietnamese Part-of-Speech Tagging Tools. 2020.

[160] Nguyen B D, Nguyen K V, Nguyen L T. LSTM Easy-first Dependency Parsing with Pre-trained Word Embeddings and Character-level Word Embeddings in Vietnamese[C]// 2018 10th International Conference on Knowledge and Systems Engineering (KSE). 2018.

[161] Thi L N, My L H, Minh H, et al. Using BiLSTM in dependency parsing for Vietnamese[J]. Computacion y Sistemas, 2018, 22(3):853-862.

[162] Xianwen Liao, Yongzhong Huang, Yongzhuang Wei, Chenhao Zhang, Fu Wang, Yong Wang. Efficient Estimate of Sentence's Representation Based on the Difference Semantics Model. IEEE

Transactions on Audio, Speech & Language Processing, 29:3384-3399, 2021. [doi]

[163] Tuyen Thi-Thanh Do and Dang Tuan Nguyen. 2021. A computational semantic information retrieval model for Vietnamese texts. Int. J. Comput. Sci. Eng. 24, 3 (2021), 301–311.

[164] 徐毓,赖华,余正涛,高盛祥,文永华.基于深度可分离卷积的汉越神经机器翻译[J].厦门大学学报(自然科学版),2020,59(02):220-224.

[165] 王振晗,何建雅琳,余正涛,文永华,郭军军,高盛祥.融合句法解析树的汉-越卷积神经机器翻译[J].软件学报,2020,31(12):3797-3807.DOI:10.13328/j.cnki.jos.005889.

[166] 普浏清,余正涛,文永华,高盛祥,刘奕洋.基于依存图网络的汉越神经机器翻译方法[J].中文信息学报,2021,35(12):68-75.

[167] Jiang H, He Y, Liao M, et al. English-Vietnamese machine translation model based on sequence to sequence algorithm[C]// 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE, 2020.

[168] My, Le Hoang Thi and Khanh Phan Huy. “Basing on the Ede syllable models to check Ede syllable misspelling, applying to improve the quality of Ede vocabulary corpus.” 2016 International Conference on Advanced Technologies for Communications (ATC) (2016): 158-162.

[169] Nguyen, V.H., Nguyen, H.T., Snasel, V. (2015). Normalization of Vietnamese Tweets on Twitter. In: Abraham, A., Jiang, X., Snášel, V., Pan, JS. (eds) Intelligent Data Analysis and Applications. Advances in Intelligent Systems and Computing, vol 370. Springer, Cham.

[170] Tran O T, Bui V T. A BERT-based Hierarchical Model for Vietnamese Aspect Based Sentiment Analysis[C]// 2020 12th

International Conference on Knowledge and Systems Engineering (KSE). 2020.

[171] Phan L L, Pham P H, Nguyen T T, et al. SA2SL: From Aspect-Based Sentiment Analysis to Social Listening System for Business Intelligence[C]// 2021.

[172] Nguyen V H, Nguyen T C, Nguyen M T, et al. VNDS: A Vietnamese Dataset for Summarization[C]// 2019 6th NAFOSTED Conference on Information and Computer Science (NICS). 2019.

[173] Tran N T, Nghiem M Q, Nguyen N T H, et al. ViMs: a high-quality Vietnamese dataset for abstractive multi-document summarization[J]. Language Resources and Evaluation, 2020, 54(4):893-920.

[174] Nguyen T S, Nguyen L M. Nested Named Entity Recognition Using Multilayer Recurrent Neural Networks[M]. 2018.

[175] 孙帅强. 面向互联网新闻的汉语—泰语双语语料挖掘方法研究[D].昆明理工大学,2018.

[176] 张金鹏. 汉泰双语新闻话题发现方法研究[D].昆明理工大学,2016.

[177] Lapjaturapit T, Viriyayudhakom K, Theeramunkong T. Multi-candidate word segmentation using bi-directional LSTM neural networks[C]//2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES). IEEE, 2018: 1-6

[178] Chormai P, Prasertsom P, Rutherford A. Attacut: A fast and accurate neural thai word segmenter[J]. arXiv preprint arXiv:1911.07056, 2019.

[179] Seeha S, Bilan I, Sanchez L M, et al. ThaiLMCut: Unsupervised pretraining for Thai word segmentation[C]//Proceedings of the 12th Language Resources and Evaluation Conference. 2020: 6947-6957

[180] 陶广奉,线岩团,王红斌,汪淑娟.融合上下文字符信息的泰语神经网络分词方法[J].计算机工程与科学,2018,40(05):943-949.

[181] 吴辉文. 泰语分词与实体抽取技术研究[D].桂林电子科技大学,2021.DOI:10.27049/d.cnki.ggldc.2021.000829.

[182] Sornlertlamvanich V, Charoenporn T, Isahara H. ORCHID: Thai part-of-speech tagged corpus[J]. National Electronics and Computer Technology Center Technical Report, 1997: 5-19.

[183] 赵世瑜. 泰语词法分析关键技术研究[D].昆明理工大学,2016.

[184] 陶广奉. 基于跨语言迁移学习的泰语依存句法解析方法研究[D].昆明理工大学,2017.

[185] 洪玄贵(WUTTHITHANAKON WUTTHIPONG). 泰语句子相似度计算研究[D].昆明理工大学,2017.

[186] Thattinaphanich S, Prom-on S. Thai named entity recognition using Bi-LSTM-CRF with word and character representation[C]// 2019 4th International Conference on Information Technology (InCIT). IEEE, 2019: 149-154.

[187] Liu Y. Fine-tune BERT for extractive summarization[J]. arXiv preprint arXiv:1903.10318, 2019.

[188] Thattinaphanich S, Prom-on S. Thai named entity recognition using Bi-LSTM-CRF with word and character representation[C]// 2019 4th International Conference on Information Technology (InCIT). IEEE, 2019: 149-154.

[189] 王红斌,郜洪奎,沈强,线岩团.泰语人名、地名、机构名实体识别研究[J]. 系统仿真学报,2019,31(05):1010-1018.

DOI:10.16182/j.issn1004731x.joss.17-0163.

[190] 吴辉文. 泰语分词与实体抽取技术研究[D].桂林电子科技大学,2021.

[191] Buaphet W, Udomcharoenchaikit C, Limkonchotiwat P, et al. Thai Nested Named Entity Recognition Corpus[C]//Findings of the Association for Computational Linguistics: ACL 2022. 2022: 1473-1486

[192] Lowphansirikul L, Polpanumas C, Jantrakulchai N, et al. Wangchanberta: Pretraining transformer-based thai language models[J]. arXiv preprint arXiv:2101.09635, 2021.

[193] Bi K, Jha R, Croft W B, et al. Aredsum: Adaptive redundancy-aware iterative sentence ranking for extractive document summarization[J]. arXiv preprint arXiv:2004.06176, 2020.

[194] 李思卓(2019)基于图匹配的老-汉双语平行句对抽取方法研究，昆明理工大学

[195] 博恩(SISOUMANG BOUANGEUN).(2018).老—汉双语语料库系统构建研究(硕士学位论文,昆明理工大学).

[196] 殷若尘. 汉—老双语词语对齐及依存树库构建方法研究[D].昆明理工大学,2017.

[197] Tien H N, Huu D N, Le Thanh H, et al. KC4Align: Improving Sentence Alignment method for Low-resource Language Pairs[C]// Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation. 2021: 358-367.

[198] Qiu X., H. Xue, L. Liang, Z. Xie, S. Liao and G. Shi, "Automatic Generation of Multiple-choice Cloze-test Questions for Lao Language Learning," 2021 International Conference on Asian Language Processing (IALP), 2021, pp. 125-130

[199] Yu, Z., Yu, Z., Huang, Y., Guo, J., Wang, Z., Man, Z. (2020). Transfer Learning for Chinese-Lao Neural Machine Translation with Linguistic Similarity. In: Li, J., Way, A. (eds) Machine Translation. CCMT 2020. Communications in Computer and Information Science, vol 1328. Springer, Singapore

[200] 杨蓓,周兰江,余正涛,刘丽佳.半监督学习的老挝语词性标注方法研究[J].计算机科学,2016,43(09):103-106.

[201] 王兴金,周兰江,张建安,周枫. 融合词结构特征的多任务老挝语词性标注方法[J]. 中文信息学报, 2019, 33(11): 39-45.

[202] 王兴金, 周兰江, 张建安, & 周枫. (2019). 融合词结构特征的多任务老挝语词性标注方法. 中文信息学报, 33(11), 39-45.

[203] 彭骁男,周兰江,张建安,周枫.融合多特征的老挝语人名地名命名实体识别[J].中国水运(下半月),2020,20(03):74-77

[204] 殷若尘. 汉—老双语词语对齐及依存树库构建方法研究[D].昆明理工大学,2017.

[205] 李炫达,汉老双语文本相似度计算方法研究[D]. 昆明理工大学,2021.

[206] Zhuo Chen, Lan Jiang Zhou, Xuan Da Li, Jia Nan Zhang, and Wen Jie Huo. 2020. The Lao Text Classification Method Based on KNN. Procedia Comput. Sci. 166, C (2020), 523–528.

[207] 何阳宇,易晓宇,唐亮,易绵竹,李宏欣.基于 BLSTM-ATT 的老挝语军事领域实体关系抽取[J].计算机技术与发展,2021,31(05):31-37

[208] Li X. and L. Zhou, "Similarity Computing Method of Poly-Encoders Short Texts in Both Chinese and Lao Based on Part of Speech Characteristics," 2021 3rd International Conference on Applied Machine Learning (ICAML), 2021, pp. 267-271

[209] 郭雷,周兰江,周蕾越.融合词语多特征的汉老短文本相似度计

算 [J]. 小 型 微 型 计 算 机 系 统 :1-9[2022-06-10]. DOI:10.20009/j.cnki.21-1106/TP.2021-0626.

[210] 李思卓,周兰江,周枫,张建安.基于互译特征词对匹配的老-汉双语句子相似度计算方法研究[J].现代电子技术,2019, 42(24):79-83+87.DOI:10.16652/j.issn.1004-373x.2019.24.019.

[211] 谭琪辉,周兰江,刘畅. 融合文本特征的汉老双语句子相似度计算方法[J]. 中文信息学报, 2021, 35(10): 64-72.

[212] 何力. 汉老双语句子相似度计算方法研究[D]. 昆明理工大学, 2019. DOI:10.27200/d.cnki.gkmlu.2019.001236.

[213] 杨志婥琪, 周兰江, 周蕾越. 融合字形特征的多任务老挝语文字识别后纠错[J/OL]. 小型微型计算机系统: 1-9 [2022-06-10].

[214] Mangeot M. Mot project: conversion of a French-Khmer published dictionary for building a multilingual lexical system[J]. arXiv preprint arXiv:1405.5674, 2014.

[215] 刘小惠. 汉柬双语可比语料库构建方法研究[D].昆明理工大学,2016.

[216] 潘丽同. 基于 Web 的英柬双语平行句对获取[D].昆明理工大学,2015.

[217] 李思远. 基于双向循环神经网络的可比语料库柬汉平行句对获取[D].昆明理工大学, 2019. DOI:10.27200/d.cnki.gkmlu.2019.002114.

[218] Chi H, Yan X, Li S, et al. The acquisition of Khmer-Chinese parallel sentence pairs from comparable corpus based on manhattan-BiGRU model[C]//2020 Chinese Control And Decision Conference (CCDC). IEEE, 2020: 4801-4805.

[219] Suraiya Jabin, Niladri Chatterjee, Suos Samak, Kim Sokphyrum, Javier Sola. An online English-Khmer hybrid machine translation

system[J]. Int. J. of Intelligent Systems Technologies and Applications, 2018,17(3).

[220] Prasomsuk Sukchatri, Mol Puthy. Thai to Khmer Rule-Based Machine Translation Using Reordering Word to Phrase[J]. International Journal of Computer Theory and Engineering,2017,9(3).

[221] Marie B, Kaing H, Mon A M, et al. Supervised and unsupervised machine translation for Myanmar-English and Khmer-English[C]// Proceedings of the 6th Workshop on Asian Translation. 2019: 68-75.

[222] Tran Van Nam, Nguyen Thi Hue, Phan Huy Khanh. Building a Syllable Database to Solve the Problem of Khmer Word Segmentation[J]. International Journal on Natural Language Computing, 2017,6(1).

[223] 潘华山,严馨,周枫,余正涛,郭剑毅.基于层叠条件随机场的高棉语分词及词性标注方法[J].中文信息学报,2016,30(04):110-116.

[224] Buoy R, Taing N, Kor S. Khmer word segmentation using bilstm networks[C]//4th Regional Conference on OCR and NLP for ASEAN Languages. 2020.

[225] Sangvat, S., Pluempitiwiriyawej, C.: Khmer POS tagging using conditional random fields. In: Hasida, K., Pa, W.P. (eds.) PACLING 2017. CCIS, vol. 781, pp. 169–178. Springer, Singapore

[226] Sry S, Nguyen A S. A review of Khmer word segmentation and part-of-speech tagging and an experimental study using bidirectional long short-term memory[J]. HO CHI MINH CITY OPEN UNIVERSITY JOURNAL OF SCIENCE-ENGINEERING AND TECHNOLOGY, 2022, 12(1): 23-34.

[227] 黄淑慧. 基于约束条件随机场的高棉语命名实体识别研究[D]. 昆明理工大学,2016.

[228] 徐广义,严馨,余正涛,周丽华.融合跨语言特征的高棉语命名实体识别方法[J].云南大学学报(自然科学版),2018,40(05):865-871.

[229] 郭月江. 利用跨语言特征的高棉语命名实体识别研究[D].昆明理工大学,2018.

[230] 谢俊. 基于主题模型词向量的高棉语命名实体识别[D].昆明理工大学,2019.

[231] 徐璐. 高棉语依存句法分析方法研究[D].昆明理工大学,2017.

[232] Kann B, Chay-intr T, Kaing H, et al. Khmer Treebank Construction via Interactive Tree Visualization[J]. IJITEE (International Journal of Information Technology and Electrical Engineering),2019 3(3): 67-74.

[233] 李小龙(TRY RATANAK). 高棉语新闻评论文本情感分类研究[D].昆明理工大学,2017.

[234] Rifat M R I, Imran A A. Incorporating Transformer Models for Sentiment Analysis and News Classification in Khmer[C]// International Conference on Computational Data and Social Networks. Springer, Cham, 2021: 106-117.

[235] 毛存礼,吴霞,朱俊国,余正涛,李云龙,王振晗.基于 CNN-CorrNet 网络的汉缅平行句对抽取方法[J].中文信息学报,2020,34(11):60-66.

[236] 李训宇,毛存礼,余正涛,高盛祥,王振晗,张亚飞.融合主题模型及双语词向量的汉缅双语可比文档获取方法[J].中文信息学报,2021,35(01):88-95.

[237] 毛存礼,高旭,余正涛,王振晗,高盛祥,满志博.结构特征一致性约束的双语平行句对抽取[J].重庆大学学报,2021,44(01):46-56.

[238] 毛存礼,陆杉,王红斌,余正涛,吴霞,王振晗.基于半监督的汉缅双语词典构建方法[J].中文信息学报,2021,35(07):47-53.

[239] 李越,毛存礼,余正涛,高盛祥,王振晗,张亚飞.融合主题及上下文特征的汉缅双语词汇抽取方法[J].小型微型计算机系统,2021,42(01):91-95.

[240] 林颂凯,毛存礼,余正涛,郭剑毅,王红斌,张家富.基于卷积神经网络的缅甸语分词方法[J].中文信息学报,2018,32(06):62-70+79.

[241] 马昌娥, 杨鉴. 缅甸语分词方法及其实现[J]. 计算机科学与应用, 2018, 8(11): 1682-1688. https://doi.org/10.12677/CSA.2018.811185

[242] Minn K H, Soe K M. Myanmar Word Stemming and Part-of-Speech Tagging using Rule Based Approach[D]. MERAL Portal, 2019.

[243] Cing D L, Soe K M. Improving accuracy of part-of-speech (POS) tagging using hidden markov model and morphological analysis for Myanmar Language[J]. International Journal of Electrical and Computer Engineering, 2020, 10(2): 2023.

[244] C. Ding, Y. K. Thu, M. Utiyama, and E. Sumita, “Parsing Myanmar (Burmese) by Using Japanese as a Pivot”, Proceedings of 14th International Conference on Computer Applications, 158-162, 2016.

[245] Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages", 5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU Workshop), 09-12 May, 2016, Yogyakarta, Indonesia, Procedia Computer Science, Volume 81, 2016, pp. 250–257.

[246] Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, “A Large-scale Study of Statistical Machine Translation Methods for

Myanmar Language", in Proc. of SNLP2016, February 10-12, 2016.

[247] Sin Y M S. Syllable-based Neural Machine Translation System for Myanmar-English Language Pair[D]. MERAL Portal, 2019.

[248] 满志博,毛存礼,余正涛,李训宇,高盛祥,朱俊国.基于多语言联合训练的汉-英-缅神经机器翻译方法[J].清华大学学报(自然科学版),2021,61(09):927-935.

[249] "Sentiment Analysis of Review of Restaurant in Myanmar Text", June 26-28, 2017, Kanazawa, Japan, 2017 IEEE SNPD 2017.

[250] Maw S Y, Khine M A. Aspect based Sentiment Analysis for travel and tourism in Myanmar Language using LSTM[D]. MERAL Portal, 2019.

[251] Aung H M S, Pa W P. Analysis of word vector representation techniques with machine-learning classifiers for sentiment analysis of Public Facebook Page's Comments in Myanmar Text[C]//2020 IEEE Conference on Computer Applications (ICCA). IEEE, 2020: 1-7.

[252] Phyu M S, Nwet K T. A study on a joint deep learning model for Myanmar text classification[C]//2020 IEEE Conference on Computer Applications (ICCA). IEEE, 2020: 1-4.

[253] Thu Y, Pa W P. Generating Myanmar News Headlines using Recursive Neural Network[C]//2020 IEEE Conference on Computer Applications (ICCA). IEEE, 2020: 1-6.

[254] Mon A M, Soe K M. Clustering analogous words in myanmar language using word embedding model[D]. MERAL Portal, 2019.

[255] Mon A M, Ding C, Kaing H, et al. A Myanmar (Burmese)-English named entity transliteration dictionary[C]//Proceedings of the 12th Language Resources and Evaluation Conference. 2020: 2980-2983.

[256] 刘梦媛,杨鉴.基于HMM的缅甸语语音合成系统设计与实现[J].

云南大学学报(自然科学版),2020,42(01):19-27.

[257] 杨馨,杨鉴. 面向文语转换的缅语音节划分和文本罗马化[A].中国中文信息学会语音信息专业委员会.第十四届全国人机语音通讯学术会议(NCMMSC’2017)论文集[C].中国中文信息学会语音信息专业委员会:清华信息科学与技术国家实验室(筹),2017:6.

[258] 毛存礼,谢旭阳,余正涛,高盛祥,王振晗,刘福浩.基于知识蒸馏的缅甸语光学字符识别方法[J].数据采集与处理,2022,37(01):173-182.DOI:10.16337/j.10049037.2022. 01.015.

[259] 黄水清，王东波.国内语料库研究综述[J].信息资源管理学报，2021，11(03):4-17+87.

[260] 赵斯琴，高光来，何敏.蒙古语语料库的研究与建设[J].内蒙古大学学报(自然科学版)，2003(05):578-581.

[261] 淑琴，那顺乌日图.面向 EBMT 系统的汉蒙双语语料库的构建[J].内蒙古社会科学(汉文版)，2006(01):140-144.

[262] 赵栋材.面向藏语自然语言处理的藏语语言资源建设[J].西藏科技，2012(09):74-77.

[263] 陈小莹，艾金勇.近十年我国藏文信息研究的特征分布与热点分析——基于 CNKI 核心期刊的文献计量及可视化分析[J].西藏民族大学学报(哲学社会科学版)，2020，41(03):141-147.

[264] 高定国，扎西加，赵栋材.“大型藏文基础语料库”数据分析[J].西北民族大学学报(自然科学版)，2013，34(04):46-51.

[265] 高定国，杨晓龙，杨宇帆，取次，高红梅.MLWS2021 藏文分词评测报告[J].高原科学研究，2022，6(01):82-89.

[266] 才让加. 藏语语料库词语分类体系及标记集研究[J]. 中文信息学报， 2009， 23(4):6.

[267] 康才畯,龙从军,江荻.基于条件随机场的藏文人名识别研究[J].计算机工程与应用，2015，51(03):109-111+185.

[268] 李博涵，刘汇丹，龙从军，吴健.基于深度学习的藏文分词方法[J].计算机工程与设计，2018，39(01):194-198.才让加.面向自然语言处理的大规模汉藏(藏汉)双语语料库构建技术研究[J].中文信息学报，2011，25(06):157-161.

[269] 龙从军，刘汇丹，周毛克.基于句法树的藏语最长名词短语识别[J].中文信息学报，2019，33(02):59-66.

[270] 华却才让，姜文斌，赵海兴，刘群.基于词对依存分类的藏语树库半自动构建研究[J].中文信息学报，2013，27(05):166-172.

[271] 扎西加，多拉.藏语依存树库构建的理论与方法探析[J].西藏大学学报(自然科学版)，2015，30(02):76-83.

[272] 夏吾吉，黄鹤鸣，华却才让.基于语义关系的藏语依存树库构建研究[J].电子技术与软件工程，2021(20):128-130.

[273] 多杰卓玛. 藏语语义框架的理解与描述[J]. 西北民族大学学报（自然科学版）， 2009(2):6.

[274] 才让三智，多拉.面向信息处理的藏语虚词知识库构建研究[J].西北民族大学学报(自然科学版)，2012，33(02):40-43+80.

[275] 祁坤钰.面向信息处理的藏语语义角色研究[J].西北民族大学学报(自然科学版)，2014，35(04):19-26.

[276] 柔特.基于 WordNet 的藏文语义词典半自动构建方法研究[J].西藏大学学报(自然科学版)，2014，29(01):48-53.

[277] 龙从军，周毛克，刘汇丹.基于词向量的藏文语义相似词知识库构建[J].中文信息学报，2020，34(10):33-38+50.

[278] 杨欣，群诺，郭龙银，孟姚媛.藏文情感语料库的构建与分析[J].计算机时代，2019(09):5-7+12.

[279] 哈里旦木·阿布都克里木，孙茂松，刘洋，阿布都克力木·阿布力孜.THUUyMorph:维吾尔语形态切分语料库[J].中文信息学报，2018，32(02):81-86.

[280] 阿西穆•托合提，早克热•卡德尔，吐尔根•依布拉音，艾山•吾买尔.乌兹别克语-维吾尔语双语语料库构建平台的设计与实现[J].电脑知识与技术，2017，13(07):1-2+10.

[281] 冯韬，李淼，曹宜超，曾伟辉.汉维可比语料数据集[J].中国科学数据(中英文网络版)，2020，5(01):167-172.

[282] 伊尔夏提•吐尔贡，吾守尔•斯拉木，热西旦木•吐尔洪太，于清.维吾尔文情感语料库的构建与分析[J].计算机与现代化，2017(04):67-72.

[283] 年梅，范祖奎，刘若兰.维吾尔语褒贬情感词典构建研究[J].计算机工程与应用，2017，53(04):152-155+162.

[284] 王成平.信息处理用彝、汉、英三语平行语料库的建设与语料对齐技术研究[J].科技通报，2012，28(02):131-133.

[285] 刘娟.内蒙古旅游与外宣资料蒙汉英三语平行语料库建设的构想[J].内蒙古师范大学学报(哲学社会科学版)，2016，45(05):160-163.

[286] 张羽.壮族嘹歌壮-汉-英三语平行语料库构建及其应用[J].百色学院学报，2016，29(01):127-130.

[287] 周秀苗.壮族典籍多语平行语料库建设与应用研究[J].社科纵横，2017，32(10):137-139.

[288] 董青秀，穗志方，詹卫东等.自然语言处理评测中的问题与对策[J].中文信息学报，2021，35(6):1-15.

[289] ZHANG N，CHEN M，BI Z 等.CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark: arXiv: 2106.08087[R/OL].arXiv，2022[2022-06-14].

[290] DUAN X，WANG B，WANG Z 等.CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension[C/OL]//SUN M，HUANG X，JI H 等. Chinese

Computational Linguistics. Cham: Springer International Publishing，2019:439-451.

[291] ZHANG S，ZHANG X，WANG H 等.Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection[J/OL].IEEE Access，2018，6:74061-74071.

[292] 唐琳，郭崇慧，陈静锋.中文分词技术研究综述[J].数据分析与知识发现，2020(2):17.

[293] 蒋俊杰.机器翻译评测技术研究[D].北京交通大学，2013.

[294] 张卫晴，张政.从机器翻译评测看机器翻译发展[J].中国科技翻译，2008(2):5.

[295] Han L. An Overview on Machine Translation Evaluation[J].arXiv e-prints，2022.

[296] 赵小兵，高璐，高定国，孙媛.少数民族语言分词技术评测数据集 MLWS2021.多语种智能信息处理数据集专刊.2022.1.28

[297] Jan A. Botha, Chris Dyer, and Phil Blunsom. 2012. Bayesian Language Modelling of German Compounds. In Proceedings of COLING 2012, pages 341–356, Mumbai, India. The COLING 2012 Organizing Committee.

[298] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics （Volume 1: Long Papers）. 2016: 1715-1725.

[299] Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics （Volume 1: Long Papers）. 2018: 66-75.

[300] Kudo T, Richardson J. SentencePiece: A simple and language

independent subword tokenizer and detokenizer for Neural Text Processing[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018: 66-71.

[301] Kudo T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018: 66-75.

[302] Xu J, Zhou H, Gan C, et al. Vocabulary Learning via Optimal Transport for Neural Machine Translation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 7361-7373.

[303] 古丽拉·阿东别克，米吉提·阿布力米提.维吾尔语词切分方法初探[J].中文信息学报，2004（6）：61-65.

[304] 陈鹏.基于语料库的维吾尔语词干提取和词性标注[D].乌鲁木齐：新疆大学，2006.

[305] 艾山·吾买尔，吐尔根·依不拉音，早克热·卡德尔.维吾尔语名词词干提取算法的研究[C]//第四届全国信息检索与内容安全学术会议，2008：180-186.

[306] ORHUN M，TANTUG A，ADALO E.Rule based analysis of the Uyghur nouns[J].International Journal on Asian Language Processing，2008，19(1)：33-43.

[307] AISHA B ， SUN M.A statistical method for Uyghur tokenization[C]//Natural Language Processing and Knowledge Engineering，2009

[308] 艾山·吾买尔，吐尔根·依步拉音，早克热·卡德尔.基于噪

声信道的维吾尔语央音原音识别模型[J].计算机工程与应用，2010，46（15）：118-120.

[309] 麦热哈巴·艾力，姜文斌，吐尔根·依布拉音.维吾尔语词法中音变现象的自动还原模型[J].中文信息学报，2012 26（1）：91-96

[310] 张海波，蔡洽吾，姜文斌等.基于联合音变还原和形态切分的形态分析方法[J].中文信息学报，2014，28（6）：9-17

[311] 米尔阿迪力江·麦麦提.基于 Morfessor 的维吾尔语词干提取和词性标注的研究[D].乌鲁木齐：新疆大学，2015.

[312] Tursun E，Ganguly D，Osman T，et al. A semisupervised tag-transition-based Markovian model for Uyghur morphology analysis[J].ACM Transactions on Asian & Low Resource Language Information Processing， 2016，16（2）：1-23.

[313] 哈里旦木·阿布都克里木，程勇，刘洋，等.基于双向门限递归单元神经网络的维吾尔语形态切分[J].清华大学学报（自然科学版），2017，57（1）：1-6.

[314] 吐尔洪·吾司曼，杨雅婷，艾孜孜·吐尔逊等.字符级的维吾尔语形态协同分析方法[J].北京大学学报（自然科学版），2019，55（1）：47-54.

[315] ABULIMITI A，SCHULTZ T.Building language models for morphological rich low-resource languages using data from related donor languages：the case of Uyghur[C/OL]// Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages（SLTU） and Collaboration and Computing for Under- Resourced Languages（CCURL）.Marseille，France：European Language Resources Association，2020：271-276（2020-05）[2021-02- 23].https：//www.aclweb.org/anthology/2020.sltu-1.38/.

[316] Huang J H, Powers D. Chinese word segmentation based on contextual entropy[C]//Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation. 2003: 152-158.

[317] Peng F, Feng F, McCallum A. Chinese segmentation and new word detection using conditional random fields[C]//COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. 2004: 562-568.

[318] 张华平, 商建云. 面向社会媒体的开放领域新词发现[J]. 中文信息学报, 2017, 31（3）: 55-61.

[319] Luo S, Sun M. Two-character Chinese word extraction based on hybrid of internal and contextual measures[C]//Proceedings of the second SIGHAN workshop on Chinese language processing. 2003: 24-30.

[320] Huang M, Ye B, Wang Y, et al. New word detection for sentiment analysis[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics（Volume 1: Long Papers）. 2014: 531-541.

[321] McCrae J P. Identification of adjective-noun neologisms using pretrained language models[C]//Proceedings of the Joint Workshop on Multiword Expressions and WordNet （MWE-WN 2019）. 2019: 135-141.

[322] Liang Y, Yin P, Yiu S M. New word detection and tagging on Chinese Twitter stream[C]//International Conference on Big Data Analytics and Knowledge Discovery. Springer, Cham, 2015: 310-321.

[323] 张婧, 黄锴宇, 梁晨, 等. 面向中文社交媒体语料的无监督新词识别研究[J]. 中文信息学报, 2018, 32（3）: 17-25, 33.

[324] Qian Y, Du Y, Deng X, et al. Detecting new Chinese words from massive domain texts with word embedding[J]. Journal of Information Science, 2019, 45（2）: 196-211.

[325] 张乐, 冷基栋, 吕学强, 等. MWEC: 一种基于多语义词向量的中文新词发现方法[J]. 数据分析与知识发现, 2022, 6（1）: 113-121.

[326] Xie T, Wu B, Wang B. New word detection in ancient Chinese literature[C]//Asia-Pacific web （APWeb） and web-age information management （WAIM） joint conference on web and big data. Springer, Cham, 2017: 260-275.

[327] 刘昱彤, 吴斌, 谢韬, 等. 基于古汉语语料的新词发现方法[J]. 中文信息学报, 2019, 33（1）: 46-55.

[328] Humbley J. Les dictionnaires de néologismes, leur évolution depuis 1945 : une perspective européenne. 2008.

[329] Cartier E. Neoveille, a Web Platform for Neologism Tracking[C]// Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017.

[330] Uchiumi K, Tsukahara H, Mochihashi D. Inducing word and part-of-speech with Pitman-Yor hidden semi-Markov models[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing （Volume 1: Long Papers）. 2015: 1774-1782.

[331] Ingrid Falk, Delphine Bernhard, Christophe Gérard. From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. LREC - The 9th edition of the Language Resources and Evaluation Conference, May 2014, Reykjavik, Iceland.hal-00959079

[332] Klosa A, Lüngen H. New German words: detection and description[J]. 2018.

[333] Firat, O., Cho, K., Bengio, Y. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016), 866-875.

[334] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Dean, J. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics (TACL), 5:339-351.

[335] Ha, T. L., Niehues, J., Waibel, A. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In Proceedings of the 13th International Conference on Spoken Language Translation (IWSLT2016).

[336] Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Wu, Y. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 3874-3884.

[337] Kudugunta, S., Bapna, A., Caswell, I., Firat, O. 2019. Investigating Multilingual NMT Representations at Scale. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), 1565-1575.

[338] Blackwood, G., Ballesteros, M., Ward, T. 2018. Multilingual

Neural Machine Translation with Task-Specific Attention. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 3112-3122.

[339] Sachan, D., Neubig, G. 2018. Parameter Sharing Methods for Multilingual Self-Attentional Translation Models. In Proceedings of the Third Conference on Machine Translation: Research Papers (WMT 2018), 261-271.

[340] Bapna, A., Firat, O. 2019. Simple, Scalable Adaptation for Neural Machine Translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), 1538-1548.

[341] Zaremoodi, P., Buntine, W., Haffari, G. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018) , 656-661.

[342] Platanios, E. A., Sachan, M., Neubig, G., Mitchell, T. 2018. Contextual Parameter Generation for Universal Neural Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 425-435.

[343] Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., Liu, T. Y. 2019. Multilingual neural machine translation with knowledge distillation. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2019).

[344] Aharoni, R., Johnson, M., Firat, O. 2019. Massively Multilingual Neural Machine Translation. In Proceedings of the 2019 Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 3874-3884.

[345] Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., Sun, J. 2018. A neural interlingua for multilingual machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers (WMT 2018), 84-92.

[346] Vázquez, R., Raganato, A., Tiedemann, J., Creutz, M. 2019. Multilingual NMT with a Language-Independent Attention Bridge. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), 33-39.

[347] Murthy, R., Kunchukuttan, A., Bhattacharyya, P. 2019. Addressing word-order Divergence in Multilingual Neural Machine Translation for extremely Low Resource Languages. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 3868-3873.

[348] Wang, X., Pham, H., Dai, Z., Neubig, G. 2018. SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 856-861.

[349] Hokamp, C., Glover, J., Ghalandari, D. G. 2019. Evaluating the Supervised and Zero-shot Performance of Multi-lingual Translation Models. In Proceedings of the Fourth Conference on Machine Translation (WMT 2019), 209-217.

[350] Sennrich, R., Haddow, B., Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings

of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), 86-96.

[351] Fadaee, M., Bisazza, A., Monz, C. 2017. Data Augmentation for Low-Resource Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), 567-573.

[352] Wang, Y., Zhang, J., Zhai, F., Xu, J., Zong, C. 2018. Three strategies to improve one-to-many multilingual translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 2955-2960.

[353] Han, X., Zhang, Z., Zhu, J. 2021. Pre-trained Models: Past, Present and Future. AI Open, 2:225-250.

[354] Radford, A., Narasimhan, K. 2018. Improving language understanding by generative pretraining. OpenAI Blog.

[355] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), 4171–4186.

[356] Zoph, B., Yuret, D., May, J., Knight, K. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), 1568-1575.

[357] Gu, J., Wang, Y., Chen, Y., Li, V., Cho, K. 2018. Meta-Learning for Low-Resource Neural Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 3622-3631.

[358] Ren, S., Chen, W., Liu, S., Li, M., Zhou, M., Ma, S. 2018. Triangular Architecture for Rare Language Translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL2018), 56-65.

[359] Schwenk, H., Chaudhary, V., Sun, S., Gong, H., Guzmán, H. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2019), 1351-1361.

[360] Lample, G., Ott, M., Conneau, A., Denoyer, L., Ranzato, M. 2018. Phrase-Based & Neural Unsupervised Machine Translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), 5039-5049.

[361] Firat, O., Cho, K., Bengio, Y. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016), 866-875.

[362] Dabre, R., Cromieres, F., Kurohashi, S. 2017. Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. In Proceedings of Machine Translation Summit XVI: Research Track, 96-107.

[363] Garmash, E., Monz, C. 2016. Ensemble learning for multi-source neural machine translation. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers (COLING 2016), 1409-1418.

[364] Nishimura, Y., Sudoh, K., Neubig, G., Nakamura, S. 2018.

Multi-source neural machine translation with data augmentation. In Proceedings of the 15th International Conference on Spoken Language Translation (IWSLT 2018), 48-53.

[365] Nishimura, Y., Sudoh, K., Neubig, G., Nakamura, S. 2020. Multi-source neural machine translation with missing data. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(12), 569-580.

[366] 马箭飞,梁宇,吴应辉,马佳楠.国际中文教育教学资源建设 70 年：成就与展望[J].天津师范大学学报(社会科学版),2021(06):15-22.

[367] 项国雄.从传统教材到电子教材[J].信息技术教育,2005,(5).

[368] 方寅.汉语国际教育教材全媒体出版探析[J].中国出版,2013(23):42-44.

[369] 王飙.中国大陆对外汉语视听教材评述与展望[J].世界汉语教学,2009,23(02):252-261.

[370] 吴双.多媒体辅助对外汉语写作教学的意义[J].云南师范大学学报(对外汉语教学与研究版),2009,7(01):41-46.

[371] 汝淑媛.语境理论与多媒体对外汉语语法教学——以简单趋向补语的教学为例[J].中国电化教育,2011(04):113-115.

[372] 陈新.基于多模态理论框架的汉语视听说教学模式设计与研究[J].云南大学学报(自然科学版),2020,42(S1):116-122.

[373] Liu, HC. Using Eye-Tracking Technology to Explore the Impact of Instructional Multimedia on CFL Learners' Chinese Character Recognition.ASIA-PACIFIC EDUCATION RESEARCHER. 2021,30(1):33-46.

[374] 郭晶,吴应辉,谷陵,周雳,侬斐,马佳楠,崔佳兴,董晓艳.国际中文教育数字资源建设现状与展望[J].国际汉语教学研究,2021(04):86-96.

[375] 武法提,牟智佳.交互式电子教材写作工具的关键技术与基础技术框架[J].中国电化教育,2015(04):61-67.

[376] 焦燕. 基于增强现实技术的对外汉语立体化教材建设初探[C]//.第十一届中文教学现代化国际研讨会论文集.,2018:361-367.

[377] 刘英林.《国际中文教育中文水平等级标准》语法等级大纲研制及应用的若干问题——《国际中文教育中文水平等级标准·语法学习手册》前言[J].国际汉语教学研究,2022(01):54-56.

[378] Allen, D; Divekar, RR; et al. The Rensselaer Mandarin Project - A Cognitive and Immersive Language Learning Environment. THIRTY-THIRD AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE / THIRTY-FIRST INNOVATIVE APPLICATIONS OF ARTIFICIAL INTELLIGENCE CONFERENCE / NINTH AAAI SYMPOSIUM ON EDUCATIONAL ADVANCES IN ARTIFICIAL INTELLIGENCE, 2019: 9845-9846.

[379] 周晓军,马君. 一个基于 VRML 的对外汉语 E-Learning 场景设计[C]//.数字化对外汉语教学理论与方法研究.,2004:256-260.

[380] 周晓军,马君,肖静.基于 VRML 的儿童对外汉语远程教学[J].系统仿真学报,2006(S1):167-169+172.

[381] 刘哲. VRChat 在汉语国际教育中的应用研究[D].辽宁师范大学,2021.DOI:10.27212/d.cnki.glnsu.2021.001432.

[382] Cheng, K.H., & Tsai, C.C. (2012). Affordances of augmented reality in science learning: Suggestions for future research. Journal of Science Education and Technology, 22, 449-462.

[383] Wen, Y. An Augmented Paper Game With Socio-Cognitive Support, 2020, 13(2):259-268.

[384] Zhang, SL. Integrating Augmented Reality into a Task-Based Thematic Language Teaching Unit. JOURNAL OF TECHNOLOGY

AND CHINESE LANGUAGE TEACHING, 2021, 12(2):29-48.

[385] Sinyagovskaya, D; Murray, JT. Augmented Reality in Chinese Language Pronunciation Practice. 2021 IEEE INTERNATIONAL SYMPOSIUM ON MIXED AND AUGMENTED REALITY ADJUNCT PROCEEDINGS (ISMAR-ADJUNCT 2021): 403-408.

[386] 张劲松,高迎明,解焱陆.基于 DNN 的发音偏误趋势检测[J].清华大学学报(自然科学版),2016,56(11):1220-1225.

[387] 甘振业,周世华,曾浩,杨鸿武.基于 DFCNN-CTC 端到端的藏族学生普通话发音偏误检测[J].西北师范大学学报(自然科学版),2020,56(05):49-53+108.DOI:10.16783/j.cnki.nwnuz.2020.05.010.

[388] 杨龙飞,解焱陆,张劲松. 基于卷积神经网络的发音偏误趋势检测[C]//.第十四届全国人机语音通讯学术会议（NCMMSC 2017）论文集, 2017:378-382.

[389] Hu Z, Leung H, Xu Y. Automated Chinese handwriting error detection using attributed relational graph matching[C]//LNCS 5145: Advances in Web Based Learning-ICWL 2008, 2008: 344-355.

[390] Chen G, Jheng Y, Lin L. Computer-based assessment for the stroke order of Chinese characters writing[C]//Proceedings of the 2nd International Conference on Innovative Computing, Information and Control, 2007.

[391] Tang W, Leong H, Ngai G, et al. Detecting handwriting errors with visual feedback in early childhood for Chinese characters[C]//Proceedings of the 2014 Conference on Interaction Design and Children, 2014:273-276

[392] 荀恩东, 吕晓晨, 安维华等. 面向书写教学的手写汉字图像笔画还原[J]. 北京大学学报（自然科学版）, 2015, 51(2):241-248.

[393] An W, Li C. Automatic matching of character strokes for computer- aided Chinese handwriting education[C]//Proceedings of the International Conference on E-Education, Entertainment and E-Management, 2011: 283-288.

[394] 吴嘉伟.计算机辅助汉字书写教学的交互技术及关键算法研究[D].北京：北京语言大学，2017.

[395] 马乐慧. 汉字书写质量的事后评判与反馈技术研究[D]. 北京:北京语言大学, 2018.

[396] 徐品香.基于社会性网络的对外汉语教学 IAST-A 模型[J].中国电化教育,2013(07):20-24.

[397] 陈珂忆,辜亿珈. 中文社交媒体在对外汉语修辞教学中的运用及影响——以微博、微信、哔哩哔哩视频网站为例[C]//.第十一届中文教学现代化国际研讨会论文集.,2018:409-417.

[398] 李代鹏.可提供性理论视阈下基于社交网络的汉语学习研究[J].教学研究,2019,42(02):60-66+81.

[399] 谢静.基于社交媒体建构对外汉语生态课堂实效性调查研究[J].当代教育理论与实践,2022,14(02):43-49.

[400] 潘文斌. 任务型教学法在社交平台上的实践[D].复旦大学,2013.

[401] 刘丽莎. 基于社交网络平台的海外汉语文化教学方式探索[D].广东外语外贸大学,2016.

[402] Veronika Seroshtan. 基于 Instagram 汉语教学中的视觉辅助工具的设计与应用[D].华东师范大学,2017.

[403] 郑艳群.教学分析与教学计算:大数据时代汉语教学研究方法探新[J].国际汉语教学研究,2020(02):32-39.

[404] 张蕊,郑艳群.汉语阅读教学中图式理论应用形式考察与分析[J].海外华文教育,2020(01):11-18.

[405] 郑艳群,田晋华.汉语听力教学结构和过程理论模型研究[J].对外汉语研究,2020(02):114-126.

[406] 郑艳群,陆凯英.初级汉语口语课教学结构和过程理论模型研究[J].云南师范大学学报(对外汉语教学与研究版),2020,18(05):33-40.

[407] 郑艳群,周梦圆.汉语写作教学结构和过程理论模型研究[J].华文教学与研究,2020(03):37-46+54.

[408] 郑艳群,朱世芳.基础汉语综合课教学结构和过程理论模型研究[J].汉语学习,2020(01):76-83.

[409] 谢小庆. 网上模拟 HSK 考试系统和练习系统[J]. 谢小庆教育测量学论文集.

[410] 谢小庆. 网上模拟 HSK 考试系统和练习系统[J]. 谢小庆教育测量学论文集.

[411] 柴省三.计算机自适应性语言测试的智能选题方法研究[J].中国教育信息化,2014(08):81-85.

[412] Zhou, W ; Hu, RF ; et al. An Intelligent Testing Strategy for Vocabulary Assessment of Chinese Second Language Learners. INNOVATIVE USE OF NLP FOR BUILDING EDUCATIONAL APPLICATIONS, 2019: 21-29.

[413] 李琳,董璐璐,马洪超.基于 BERT 的汉语作文自动评分研究[J].中国考试,2022(05):73-80.

[414] Lee, LH; Hung, MC; et al. Chinese Grammatical Error Detection Using Adversarial ELECTRA Transformers. 29TH INTERNATIONAL CONFERENCE ON COMPUTERS IN EDUCATION (ICCE 2021), VOL I, 2021: 111-113.

[415] 韩杨超. 基于管道方式的对外汉语语法偏误自动诊断研究_韩杨超[D]. 郑州大学, 2021.

[416] 夏历,翟根广.留学生“例如”“比如”的使用情况及偏误研究.

华文教学与研究,2021(2):88-95.

[417] 李治平,李丛.HSK动态作文语料库语篇关联语使用情况统计分析.语言文字应用,2017(2):102-109.

[418] 周睿.汉日强程度副词二语习得对比研究[J].汉语学习,2021(03):94-103.

[419] 张江丽.汉语二语学习者与母语学习者产出性词汇量对比研究[J].语言文字应用,2019(02):124-132.

[420] 林柱,周小兵,郝伟.汉语集合量词二语习得研究[J].东北师大学报(哲学社会科学版),2022(03):80-87.

[421] 刘竹林.CSL学习者词语跨类混淆的分布特征与认知动因[J].河南大学学报(社会科学版),2022,62(01):112-116.

[422] 谭晓平.“A 比 B+更/还/都/再+W”的习得研究[J].汉语学习,2022(02):86-95.

[423] 郝瑜鑫,王雪琳,刘海涛.基于句法标注语料库的汉语中介语动词配价发展计量研究.语言文字应用,2021(1):29-41.

[424] 刘华,李晓源.基于语料库的中医汉语主题词表构建.华文教学与研究.2022(2):77-85.

[425] 周小兵,薄巍,王乐,李亚楠.国际汉语教材语料库的建设与应用[J].语言文字应用,2017(01):125-135.

[426] 王敬,杨丽姣,蒋宏飞,苏靖杰,付静玲.汉语二语教学领域词义标注语料库的研究及构建[J].中文信息学报,2017,31(01):221-229.

[427] 王治敏,杨尔弘.面向汉语教学的常用动词计量研究[J].语言教学与研究,2012(01):1-6.

[428] 张引兵,宋继华,彭炜明,郭冬冬,张婴,宋天宝.基于动态语料的分级词表动态生成[J].吉林大学学报(工学版),2020,50(06):2212-2220.

[429] 王贵荣,饶高琦,荀恩东.基于大规模语料库的现代汉语动宾搭配知识库构建.中文信息学报,2021,35(1):34-42.

[430] 张宝林,崔希亮.谈汉语中介语语料库的建设标准.语言文字应用,2015,(2):125-134.

[431] 黄友.面向二语学习者的汉语易混淆词语词典和语料库建设.辞书研究,2014(5):36-41.

[432] 胡晓清.国别化汉语中介语动态语料库建设理念、实践与前瞻[J].山东师范大学学报(人文社会科学版),2018,63(05):134-156.

[433] 张宝林,崔希亮.“全球汉语中介语语料库”的特点与功能.世界汉语教学,2022,36(1):90-100.

[434] 张宝林.“HSK 动态作文语料库 2.0 版”的设计理念与功能.语料库语言学,2021,8(1):81-96.

[435] 金檀,陆小飞,林筠,李百川. (2018).“汉语阅读分级指难针”.广州：语言数据网(languagedata.net/editor).

[436] 詹卫东,郭锐,常宝宝,谌贻荣,陈龙.北京大学 CCL 语料库的研制.语料库语言学,2019,6(1):71-86.

[437] 荀恩东,饶高琦,肖晓悦,臧娇娇.大数据背景下 BCC 语料库的研制.语料库语言学,2016,3(1):93-109.

[438] 全球中文学习平台：https://www.chinese-learning.cn/#/web

[439] 中文联盟：https://www.chineseplus.net/

[440] 唐风汉语国际教育平台：https://info.tangce.cn/common/index.action

[441] 长城汉语智慧云平台：https://www.greatwallchinese.com/page/qt/index.html#/wisdomPlatform

[442] Ponddy Reader：https://reader.iponddy.com/

[443] 哈兔中文网络学院：http://www.ihatoo.com/

[444] 锦灵中文：https://www.jinglelingo.net/

[445] 悟空中文：https://www.wukongsch.com/

[446] Lingo Ace：https://www.lingoace.com/cn/Home/

[447] Lingo Bus：https://m.lingobus.com/

[448] PPtutor：https://www.pptutor.com/

[449] Chinlingo：https://www.chinlingo.com/

[450] 信息来源于中外语言合作交流中心官网：

http://www.chinese.cn/uploads/file/20220125-1643091053961452.pdf