

基于Bi-LSTM和CRF的蒙古语 领域术语识别研究*

塔娜,松云

(内蒙古师范大学计算机科学技术学院,呼和浩特 010022)

摘要:针对蒙古语领域术语识别中形态复杂和语料稀缺等问题,提出融合双向长短期记忆网络(Bi-LSTM)和条件随机场(CRF)的序列标注模型Bi-LSTM+CRF。聚焦政策名词、组织机构和头衔名称3类术语,构建人工标注的蒙古语政务文本数据集,并采用BMEIO标注方法进行序列标注。通过Bi-LSTM提取上下文语义特征,结合CRF优化标签转移约束,有效应对蒙古语无大小写、复合词多和形态变化丰富等问题。实验结果表明,该模型在测试集上整体 F_1 值达93.17%,性能优于单一CRF或Bi-LSTM模型,能够有效提升蒙古语术语识别准确率,为资源稀缺语言的自然语言处理研究提供可行路径。

关键词:自然语言处理;蒙古语;领域术语;术语识别

中图分类号:TP391 **文献标志码:**A

文字不仅是语言的视觉载体,更是民族文化遗产和知识体系构建的重要媒介。在信息化与智能化深度融合的背景下,蒙古语等资源稀缺语言的自然语言处理(Natural language processing, NLP)技术的发展需要充足的高质量语料资源和高效的算法模型。蒙古语作为典型的黏着语,具有无大小写区分、形态变化丰富和复合词普遍等特点,导致其术语边界模糊、识别难度大。术语作为专业领域概念的语言指称^[1],在政务、法律和教育等场景中具有高度规范性和语义重要性,其自动识别是实现机器翻译、信息抽取和知识图谱构建等下游任务的基础。

当前,蒙古语自然语言处理仍处于发展阶段,高质量标注数据匮乏和语言结构复杂等问题严重制约了蒙古语领域术语识别技术的性能提升。为此,聚焦蒙古语政务领域,针对政策名词、组织机构与头衔名称3类关键术语,构建人工标注数据集,并设计基于深度学习的识别模型,旨在提升术语识别(Terminology recognition, TR)的准确性与鲁棒性,为蒙古语NLP的基础研究与实际应用提供技术支持。

1 相关工作

术语识别作为一项关键的信息提取方法,能够有效捕获文本中的特定名词性术语,对于推动信息检索效能提升、构建高质量知识图谱以及实现深层次知识挖掘起到重要作用。术语识别技术的发展在很大程度上受到了命名实体识别(Named entity recognition, NER)研究的启发,后者首次被正式

* 收稿日期:2025-11-27; 修回日期:2026-01-31

基金项目:内蒙古自然科学基金项目(2025MS06015)

作者简介:塔娜(1998—),女(蒙古族),内蒙古兴安盟人,2022级硕士研究生。E-mail:nomiintan@foxmail.com

通信作者:松云(1973—),女(蒙古族),内蒙古呼和浩特人,教授,博士。主要从事蒙古文信息处理、多语种教育智能化研究。E-mail:15049116307@163.com

提出是在第六届消息理解会议(MUC-6)之后^[2]。随着对文本中专业词汇和领域特定表达理解需求的增长,术语识别作为一项专门的技术获得广泛关注。

与NER类似,TR涵盖3种方法:基于规则的方法、基于统计机器学习的方法和基于深度学习的方法^[2],这些方法具有不同的技术路径和理论基础。

1.1 基于规则的方法

早期的术语识别依托专业领域的语言学规则,根据专家制定的规则库筛选与之匹配的术语。后续研究引入了浅层语法分析方法,首先依据规则库提取文本中的最长名词短语,然后通过解析这些短语的语法结构和位置特征来抽取内部的术语单元^[3]。还有研究利用单词拼写特征和词法信息抽取核心术语^[4]、采用互信息方法生成候选词^[5],在构造领域术语词表时增加语义规则与关联规则校验,优化术语的抽取效果。

基于规则的方法因其高度的灵活性而在特定领域和语言环境中展现出了独特的优势。这种灵活性体现在规则的定制化方面,即可以根据具体需求轻松地添加、修改或删除规则。规则是由人类专家定义的,因此对于那些具备专业知识背景的研究人员而言,这一过程相对透明且易于理解与调整。然而,该方法在处理复杂任务时具有一定局限性,不仅需要制定大量规则,而且这些规则的创建与维护也需耗费大量人力成本。此外,规则的质量直接取决于制定者的经验与技能水平,这在一定程度上影响了识别的准确性和可靠性。

1.2 基于统计机器学习的方法

基于统计机器学习的方法主要包括有监督和无监督两种。有监督学习方法通过从带标签的数据中学习输入与输出之间的映射关系来构建预测模型,主要包括循环神经网络(Recurrent neural network, RNN)、隐马尔可夫模型(Hidden Markov model, HMM)、支持向量机(Support vector machine, SVM)、最大熵(Maximum entropy model, MEM)和条件随机场(Conditional random field, CRF)等模型。

基于统计机器学习方法中的有监督的方法,Graves等^[6]通过训练深度RNN,在传统RNN的基础上增加了多层表示和双向学习机制,有效解决了长程依赖问题。Ma等^[7]通过结合双向LSTM、CNN和CRF模型,实现了高效的端到端序列标注,简化了特征构建和数据预处理,提升了模型性能。李丽双等^[8]通过在词频分布公式中融入信息熵,获取术语的完整语义与领域特征,并利用规则对结果进行筛选,最终获得了有效的术语。在蒙古语术语自动识别研究中,文峰^[9]根据政治学、术语学及蒙古语相关学科知识,定义了蒙古语政治术语^[9],并基于CRF模型开展自动识别研究, F_1 值达到78.40%。在其他少数民族语言领域术语识别研究方面,包括基于CRF的维吾尔语领域术语识别^[10]、基于最大熵对哈萨克语基本动词短语的识别^[11]以及基于CRF的哈萨克语领域术语识别^[12-13]。此外,在社交媒体领域^[14]与军事领域^[15]中亦存在丰富的研究成果。而针对无监督的学习方法,则是通过加入TF-IDF^[16-17]等方式,在多种类型的数据上提取领域术语。

1.3 基于深度学习的方法

近年来,基于深度学习的术语识别方法取得了显著进展,这类方法通过自动学习特征表示,降低了对人工规则和特征工程的依赖。常用模型包括递归神经网络(Recursive neural network, RNN)、双向长短期记忆网络(Bi-directional long short-term memory, Bi-LSTM)、卷积神经网络(Convolutional neural network, CNN)和Transformer架构(如BERT和RoBERTa等预训练语言模型)。

2016年,Sahu等^[18]首次在美国国家生物技术信息中心(National center for biotechnology information, NCBI)疾病数据集上使用Bi-LSTM模型捕获词序列的语义信息,并利用CNN捕获每个词的字符序列语义信息,用于识别疾病名称。Wu等^[19]采用最小特征工程方法在我国临床病历文本中进行临床实体抽取,使用两个深度神经网络(DNN)分别生成词嵌入和抽取医学实体,实验结果显示该方法优于传统的CRF方法,并证明了无监督特征学习的有效性。后期,Bi-LSTM和CRF等算法^[20-23]被成功应用于领域术语识别研究,还有不少研究者采用预训练语言模型^[24-25]进行建模。

2 蒙古语领域术语

2.1 蒙古语术语类型

蒙古语领域术语提取源自蒙古语的新闻文本数据,主要聚焦政策名词、组织机构、头衔名称3个类型。识别这3类术语,对于提升汉蒙或蒙汉机器翻译的质量至关重要,能显著增强翻译的准确性。在NLP技术方面,准确处理特定领域的术语能够提高信息检索和文本分类等任务的性能,特别是在多语言环境中,能够使NLP模型更加智能高效。

政策名词术语识别主要聚焦于能够体现思想政治教育内涵和与蒙古族文化及社会发展密切相关的专业词汇概念。如体现社会主义核心价值观、民族团结、文化认同的“*ᠪᠣᠯᠡᠭᠦᠨ ᠤ᠋ᠮᠤᠴᠢᠰᠤ ᠶᠤᠬᠤᠭᠠᠨ ᠲᠤᠳᠤᠨ ᠵᠤᠨ ᠲᠤᠩᠭ᠎ᠠ ᠵᠢᠨᠠᠨᠢ*”(中华民族共同体意识)和“*ᠰᠠᠨᠠᠨ ᠠᠨᠠᠨ ᠠᠨ ᠰᠤᠷᠭᠡᠨ ᠲᠤᠳᠤᠨ ᠵᠤᠨ ᠲᠤᠩᠭ᠎ᠠ*”(社会主义核心价值观)等;涉及经济、社会发展的战略规划、项目名称和指标体系,如“*ᠠᠨᠢ ᠪᠠᠨᠢ ᠠᠨᠢ ᠵᠤᠨ*”(一带一路)、“*ᠭᠢᠸᠠᠨᠠᠨ ᠲᠤᠨᠢ ᠵᠤᠨᠠᠨᠢ*”(绿色发展)和“*ᠵᠢᠨᠠᠨ ᠵᠤᠨᠠᠨ ᠵᠤᠨᠠᠨᠢ ᠵᠤᠨᠠᠨᠢ*”(乡村振兴)等。

组织机构术语是指蒙古语文本中出现的各种组织和机构名称,包括政府机构,如“*ᠤᠯᠤᠰ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ*”(教育厅)和“*ᠤᠯᠤᠰ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ*”(科技厅);教育组织机构,如“*ᠤᠯᠤᠰ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ*”(内蒙古大学);商业企业,如“*ᠤᠯᠤᠰ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ*”(新东方教育科技集团)。还包括非政府组织与慈善机构、媒体机构等。此外,将“*ᠤᠯᠤᠰ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ*”(中央脱贫攻坚专项工作领导小组会议)和“*ᠤᠯᠤᠰ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ ᠤᠨ ᠠᠨᠠᠭ*”(全区民用防空大会)等有明确参与者(如中央委员)、议程以及会议目标的临时性组织单元视为临时组织进行识别。

头衔名称术语可以根据不同的领域和背景,聚焦于政治头衔、宗教头衔、学术头衔、商业界头衔、军事界头衔和社会团体头衔等多个方面。

2.2 蒙古语术语特点

政策名词术语具有多重特性：①专业性和多样性，如“*མགྲོན་པོ་བྱིས་པ་ལྟར་དུ་ཕྱེད་པའི་ཁྱེད་ཀྱི་ཆུང་ཚོ་ལ་ཐུག་ཅིག་ཞུ་བཤིང་།*”(供给侧结构性改革)等术语依托专业理论，具有固定规范内涵，覆盖经济、民生、生态等多个治理领域，表述形式丰富多元；②时效性，如“*ལུས་ཅིག་གི་མཉམ་སྦྲེལ་ལྷན་ཁག་*”(计划生育)随着政策环境的变化，相应的政策名词也会不断更新，政策本身及其具体要求和执行力度都会随着时间的推移而发生显著变化；③指向性，如“*ཇ་རྒྱུ་ལ་ལཱ་རྒྱུ་ལ་ཁྱེད་ཀྱི་ལྷན་ཁག་གི་ལྷན་ཁག་གི་ལྷན་ཁག་གི་ལྷན་ཁག་གི་ལྷན་ཁག་*”(军民融合发展战略)是统筹经济建设和国防建设协调发展的一种具有明确指向性和长远规划性质的政策名词，通常直接关联政策目标、措施和执行机制，具有明确的指导意义和实施方向。

组织机构术语也呈现出多种特性:①专业性和多样性,如“ מנהל המבחנים ”(民政部)和“ מנהל המבחנים ”(政府)等具有较强的专业性,且政府组织种类较多,涵盖了各级政府机构、各部门和行政单位的名称,也包含新兴的政府组织或机构,覆盖范围广;②地域性,如“ מנהל המבחנים ”(南京市档案馆)和“ מנהל המבחנים ”(鄂尔多斯市体育局)等往往与具体地区关联;③关联性,如“ מנהל המבחנים ”(水利部综合事业局),这类术语内部存在一定的关联和层级关系;④特征化,如“ מנהל ”(委员会)和“ מנהל ”(部)等固定后缀,可以作为识别的线索和特征之一,帮助模型提高术语识别的准确性。

头衔名称术语具有以下特性：①前后关联性，常与人名搭配构成复合实体，用于表示个体在某一组织或机构内的身份或职务；②依存性，头衔名称的正确识别往往依赖于上下文信息，有时仅凭单个词难以确定是否为头衔名称，需要结合前后文理解才能做出准确判断；③特征化，如“**ᐃᓴᑦᑐᕋᔪ**”（主任）和“**ᐊᒻᒪᕈᑦ**”（部长）等常见的职称后缀可以作为识别的关键特征之一；④多样性，如“**ᐅᖁᕆ ᐱᕏᐳᕆ ᐱᕐ ᐸᕌᕈᕆᕐᐱᕐ ᐱᕐ ᐸᕌᕈᕆᕐᐱᕐ**”（国际贸易谈判副代表）等，头衔名称种类繁多，涉及不同类型的职务、称号和身份标识。

综上所述,蒙古语领域术语具有丰富的语言特征和复杂的使用情境,给术语识别研究带来诸多困难,主要难点如下。

- (1) 无大小写区分: 蒙古语与英语不同, 首字母没有大写, 因此只能依靠术语后缀识别术语边界。
- (2) 复合型名词居多: 蒙古语中大量存在由多个词汇组合而成的复合名词, 这增加了识别难度, 模型不仅需要理解单个词汇的意义, 还需要解析这些词汇如何共同构成一个完整的术语。
- (3) 形态变化丰富: 蒙古语是一种高度屈折语言, 词语根据语法会发生形态变化(如格的变化); 同时蒙古语以词为单位竖写, 字母会因在词中上下位置不同产生不同写法, 还存在较多形同音不同的现象。这种书写和语法层面的双重形态变化, 使得同一个术语可能以多种形式出现, 增加了模型归一化的复杂性^[26]。
- (4) 多义性和上下文依赖: 一些术语在不同上下文中可能有不同的含义, 增加了识别难度。

3 蒙古语领域术语识别模型

采用Bi-LSTM-CRF模型识别蒙古语领域术语, 其结构如图1所示。

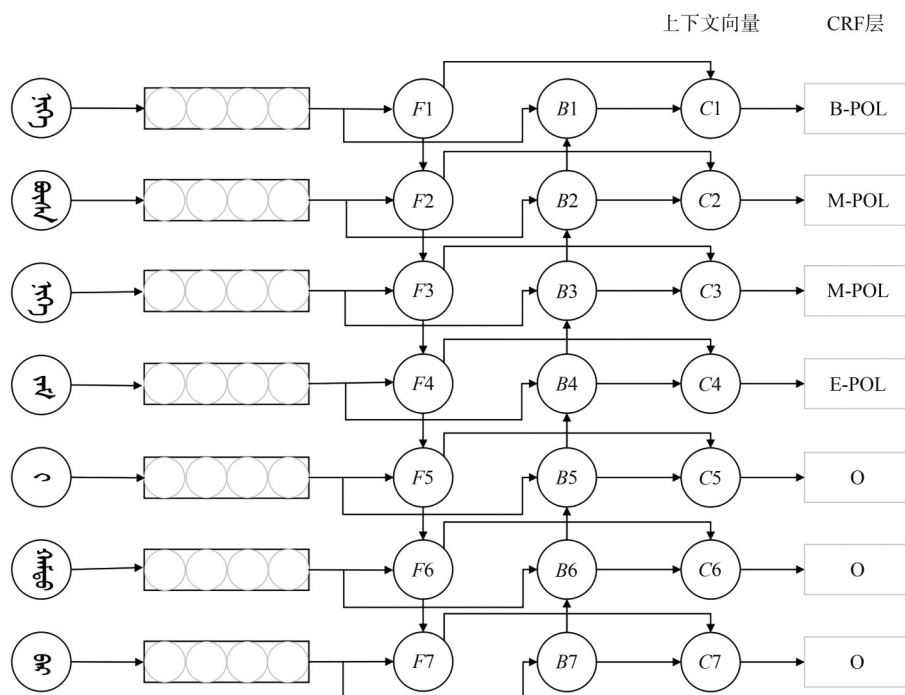


图1 Bi-LSTM-CRF模型

Fig. 1 Bi-LSTM-CRF model

3.1 双向长短期记忆网络

Bi-LSTM是基于LSTM^[27]的一种变体, 通过同时考虑序列数据的过去和未来上下文信息增强模型的特征提取能力。LSTM是RNN的一种特殊形式, 用于解决传统RNN梯度消失和梯度爆炸的问题, 通过引入输入门、遗忘门、记忆信息和输出门4个关键组件来有效控制信息流。这些门机制允许网络选择性地遗忘旧信息、存储新信息, 并结合当前输入和前一时间步的状态来更新其内部状态, 从而更好地捕捉长期依赖关系。LSTM因其出色的序列建模能力被广泛应用于命名实体识别任务^[28]。

Bi-LSTM结合了两个方向的LSTM: 一个向前传播, 一个向后传播。向前LSTM从序列的开始向结束处理信息, 能够利用过去的上下文信息; 向后LSTM则逆序处理, 专注于从序列末尾到开始的信息, 能够利用未来的上下文信息。这两个方向的输出后续被连接或合并, 以便模型能够全面地理解序列中的每一个时间点, 不仅知道“之前发生了什么”, 也了解“之后会发生什么”。LSTM的细胞结构如图2所示。

图2中符号依次为前一刻输出 $h^{(t-1)}$ 、当前输出 h^t 、当前输入 x^t 、前一刻记忆信息 $c^{(t-1)}$ 、当前记忆

信息 c' 、遗忘门 f 、输入门 i 、输出门 O , 公式分别为 $\text{sigmoid}(x)=\frac{1}{1+e^{-x}}, \tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}, i=\text{sigmoid}(w_i[x^t, h^{(t-1)}])+b_i, O=\text{sigmoid}(w_o[x^t, h^{(t-1)}])+b_o, c'=fc^{(t-1)}+i\tanh(w_c[x^t, h^{(t-1)}])+b_c, f=\text{sigmoid}(w_f[x^t, h^{(t-1)}])+b_{fo}$.

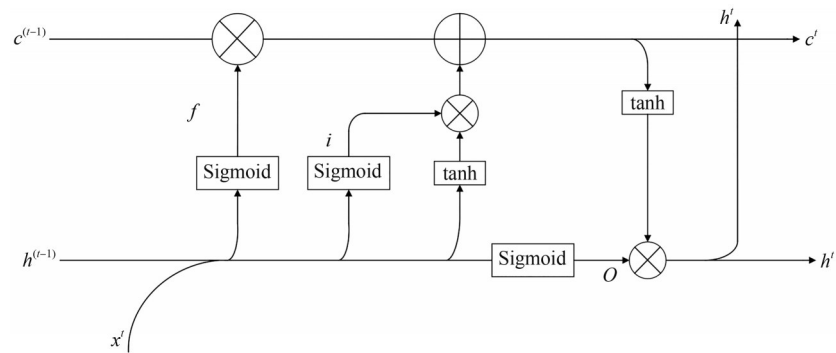


图 2 LSTM 细胞结构
Fig. 2 LSTM cell structure

3.2 条件随机场

尽管 Bi-LSTM 模型能够生成反映输入序列特性的输出,并可以直接用于标签预测,但它有时会生成不合逻辑的标签序列。为了解决这一问题,并增强模型输出的连贯性和合理性,一种常见做法是在 Bi-LSTM 后叠加 CRF 层。CRF 层能够利用序列的上下文标签信息以及标签之间的依赖关系,对 Bi-LSTM 的原始输出进行优化调整,从而提升预测的准确性和序列标签的一致性。结合 Bi-LSTM 的强大特征提取能力和 CRF 对序列结构约束的考虑,模型的预测结果能够更加准确且符合实际情境。

4 实验结果与分析

4.1 实验数据

实验数据为自建数据集,分别来自政府工作报告、政策文件和法律法规等板块。鉴于蒙古语数据的稀缺性,采用人工标注方法处理原始数据。完成标注后开展统计与分析工作,有利于更好地理解实体分布的特性、识别潜在的问题,并为后续的模式训练提供高质量的数据支持。经过统计与分析,完成数据质量校验之后,得到的术语分布如表 1 所示。

4.2 标注规则

实验中采用的实体标注格式为 BMEO。以字为单位进行标注,其中,B(Begin)代表实体类型的起始位置,M(Middle)代表实体类型的中间位置,E(End)代表实体类型的结束位置,O(Outside)代表非实体的部分。具体标注规则如表 2 所示。

表 1 术语分布 Table 1 Term distribution	
术语类型	术语规模/条
政策名词	3 056
头衔名称	5 648
组织机构	10 236

表 2 标注规则 Table 2 Annotation rules			
术语类型	开始	中间	结束
政策名词	B-POL	M-POL	E-POL
组织机构	B-ORG	M-ORG	E-ORG
头衔名称	B-TIT	M-TIT	E-TIT
非实体	O	O	O

4.3 评估指标

使用精确率(Precision, P_{rec})、召回率(Recall, R_{ec})和 F_1 值作为评估指标^[2]。精确率用来表示模型

预测正确的比例,公式为 $P_{\text{rec}}=T_p/(T_p+F_p)$;召回率用于反映模型成功找出所有实际实体的能力,公式为 $R_{\text{ec}}=T_p/(T_p+F_N)$; F_1 值是准确率和召回率的调和平均数,公式为 $F_1=2P_{\text{rec}}R_{\text{ec}}/(P_{\text{rec}}+R_{\text{ec}})$ 。其中, T_p 是模型预测为实体且实际也为实体的词数量, F_p 是模型预测为实体但实际不是实体的词数量, F_N 为模型预测不是实体但实际上为实体的词数量。

4.4 实验环境与参数设置

实验在 Windows 11 操作系统下进行,硬件配置为 NVIDIA GeForce RTX 2050 显卡。实验环境采用 Python 3.8 编程语言,搭配 PyTorch 1.12 深度学习框架。实验参数见表 3。

4.5 实验结果与分析

为检验基于 Bi-LSTM 和 CRF 的方法对蒙古语领域术语识别的可行性,将数据集按照 8:1:1 划分,进行 CRF、Bi-LSTM、Bi-LSTM+CRF 模型对比实验,结果如表 4 所示。Bi-LSTM+CRF 模型擅长捕获序列中的长程依赖性,并通过增强上下文信息的相关性,优化了模型在理解复杂语境中的能力。实验结果显示,该模型在精确率、召回率和 F_1 值上分别达到 93.05%、93.48% 和 93.17%,综合表现优异,在术语识别领域展现出良好性能。

表 3 参数设置

Table 3 Parameter settings

参数名称	参数值
batch_size	64
lr	0.01
epochs	50
print_step	5
emb_size	126
hidden_size	126

表 4 不同模型评估结果

Table 4 Evaluation results of different models

模型	P_{rec}	R_{ec}	F_1
CRF	92.84	93.45	92.84
Bi-LSTM	92.73	93.33	92.78
Bi-LSTM+CRF	93.05	93.48	93.17

各术语类别在 Bi-LSTM+CRF 模型上的测试结果如表 5 所示。从表 5 可以看出,头衔名称术语的实验效果最好,精确率达到 91.02%, F_1 值达到 84.13%, 组织机构术语的实验效果次之,政策名词术语的实验效果最差。可能原因在于:①头衔名称术语通常具有特定的构词规律和固定的表达形式,为高识别率提供了基础支撑。例如,“ $\text{ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ}$ ”(教育部部长)和“ $\text{ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ}$ ”(教育局局长)等头衔常出现在姓名之前或之后,与其所指称的个体形成语义上的紧密联系,具有明显的结构性特征。另外,头衔名称的语义边界清晰,不存在模糊歧义的表达,例如,“ ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ”(教授)和“ ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ”(主任)等头衔的指代对象是特定职业身份,使用场景以正式语境为主,不会与普通名词产生语义混淆。②组织机构术语的识别效果处于中等水平,一方面,该类术语的命名规则具有较强的明确性与固定性,为模型学习提供了清晰的特征依据;另一方面,术语的多样性、缩写等表达变异形式及与其他名词的语义相似性,共同构成了识别干扰,降低了模型的识别精度。例如,“ $\text{ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ}$ ”(中华人民共和国教育部)可缩写为“ $\text{ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ}$ ”(教育部),因缺失地域层级等核心信息,易引发语义歧义且与原称特征差异明显,对命名实体识别模型的特征匹配、歧义区分及训练学习造成干扰,降低了识别精度。③政策名词识别效果最差,原因在于其词汇形态多样、语义范围广、上下文关联复杂且专业性强。词汇形态上,既有“ $\text{ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ}$ ”(乡村振兴战略)等固定短语,也有“ $\text{ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ᠨᠠᠭᠤᠨ}$ ”(乡村新产业新业态)等新兴术语,导致模型难以归纳出稳定的判断规则。在语义层面,同一术语跨领域含义不同,如“ ᠠᠨᠠᠭᠤᠨ ᠤᠨᠢᠨᠦ ”(社会保障)在医疗、养老、就业领域指向各异。在上下文关联层面,如

“*ᠠᠵᠤᠯᠤᠰᠤ ᠠᠵᠤᠯᠤᠰᠤ ᠠᠵᠤᠯᠤᠰᠤ ᠠᠵᠤᠯᠤᠰᠤ ᠠᠵᠤᠯᠤᠰᠤ ᠠᠵᠤᠯᠤᠰᠤ*”(终身职业技能培训)需依托人社领域政策语境判定,脱离场景易与普通词汇混淆,这些因素共同增加了模型区分难度。④训练数据中政策名词的覆盖不够全面,且这些名词在文本中出现的形式多样,增加了识别难度。

表 5 Bi-LSTM+CRF 模型测试结果

Table 5 Bi-LSTM+CRF model test results

%

术语类型	P_{rec}	R_{ec}	F_1
政策名词	56.63	35.92	43.48
组织机构	76.31	74.41	74.24
头衔名称	91.02	76.14	84.13

针对实验中术语分布不均匀及各类术语识别存在的核心问题,提出以下针对性优化方案。针对头衔名称、政策名词术语样本量少及政策名词术语召回率低的问题,采用定向数据扩充策略:采集多个领域高价值文本,包括国家政策文件、行业白皮书、核心期刊论文、企业官方公告及职场沟通场景材料,确保各术语类别样本量占比控制在合理区间;结合各类术语特性设计专属特征提取方案,对于头衔名称,基于其固定构词规律,构建涵盖多领域的头衔核心词词典,设计多维度正则表达式规则,通过文本预处理的规则匹配,强化头衔名称的结构特征;对于政策名词,引入领域词典辅助特征提取,整合国家政策术语库、各行业规范词典及近几年新兴政策词汇表构建专属词表,同时采用 TF-IDF 与关键词共现网络,捕捉强关联特征,生成政策名词语义关联矩阵,辅助模型识别;对于组织机构术语,针对缩写干扰问题,构建“缩写-原称”双向映射词典,并补充缩写使用场景标签,通过词典匹配将文本中的缩写统一替换为完整原称,同时过滤各类形似、易混淆词汇,减少特征混淆。在模型层面,通过微调不同术语类别参数,如调高头衔名称结构特征权重、为政策名词引入实体级注意力机制、采用多任务学习共享 BERT 底层网络、在预训练阶段引入领域知识语料等,提升模型对各类术语的适配能力;在后处理层面,针对 3 类术语分别设计专属修正规则,如组织机构缩写还原与上下文验证、政策名词语义相似度匹配与领域标签适配、头衔名称冲突检测与修饰词兼容性判断等,优化预测结果的准确性。

5 结语

为探索一种基于蒙古语政务领域的术语识别方法,采用 Bi-LSTM+CRF 模型对蒙古语新闻文本数据进行识别实验。实验证明,Bi-LSTM+CRF 模型可以有效地识别蒙古语领域术语。通过构建更大规模、体系完善的术语语料库,融合跨语言资源,结合蒙古语语言特性优化模型,并开展持续的性能评估与迭代,蒙古语自然语言处理技术将会持续迭代升级、稳步发展。

参考文献:

[1] GB/T 10112—2019 术语工作 原则与方法[S].

[2] 康怡琳,孙璐冰,朱容波,等. 深度学习中文命名实体识别研究综述[J]. 华中科技大学学报(自然科学版),2022, 50(11):44-53.

[3] Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases[C]//Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France: Association for Computational Linguistics, 1992:361-370.

[4] Fukuda K, Tamura A, Tsunoda T, et al. Toward information extraction: Identifying protein names from biological

- papers[C]//Proceedings of the Pacific Symposium on Biocomputing. Maui, Hawaii, USA: World Scientific Publishing, 1998:707-718.
- [5] Guo J K, Van Brackle D, Lofaso N, et al. Extracting meaningful entities from human-generated tactical reports[J]. Procedia Computer Science, 2015, 61:72-79.
- [6] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013: 6645-6649.
- [7] Ma X Z, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016:1064-1074.
- [8] 李丽双,王意文,黄德根. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报, 2015, 29(1):82-87.
- [9] 文峰. 蒙古语政治术语自动识别研究[D]. 呼和浩特:内蒙古大学, 2020.
- [10] 钟军,田生伟,禹龙. Web文本中维吾尔语领域术语的自动发现[J]. 计算机应用, 2012, 32(2):407-410.
- [11] 古丽扎达·海沙. 哈萨克语基本动词短语自动识别研究[D]. 乌鲁木齐:新疆大学, 2013.
- [12] 木合亚提·尼亚孜别克,古力沙吾利·塔里甫. 哈萨克语 IT 领域术语识别研究与实现[J]. 中文信息学报, 2016, 30(3):68-73.
- [13] 木合亚提·尼亚孜别克,古力沙吾利·塔里甫,达吾勒·阿布都哈依尔. 采用CRF模型的哈萨克语信息技术术语自动抽取技术研究[J]. 西北师范大学学报(自然科学版), 2016, 52(1):53-56.
- [14] 黄鹏. 基于统一模型的中文社交媒体命名实体识别的研究[D]. 武汉:武汉邮电科学研究院, 2018.
- [15] 冯蕴天,张宏军,郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学, 2015, 42(7):15-18, 47.
- [16] 翟笃风,刘柏嵩. 政务领域本体术语的自动抽取[J]. 现代图书情报技术, 2010(4):59-65.
- [17] 闫兴龙,刘奕群,方奇,等. 基于网络资源与用户行为信息的领域术语提取[J]. 软件学报, 2013, 24(9):2089-2100.
- [18] Sahu S K, Anand A. Recurrent neural network models for disease name recognition using domain invariant features [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016:2216-2225.
- [19] Wu Y H, Jiang M, Lei J B, et al. Named entity recognition in Chinese clinical text using deep neural network[J]. Studies in Health Technology and Informatics, 2015, 216:624-628.
- [20] Dai X. Recognizing complex entity mentions: A review and future directions[C]//Proceedings of the ACL 2018 Student Research Workshop. Melbourne, Australia: Association for Computational Linguistics, 2018:37-44.
- [21] 王学锋,杨若鹏,朱巍. 基于深度学习的军事命名实体识别方法[J]. 装甲兵工程学院学报, 2018, 32(4):94-98.
- [22] 单义栋,王衡军,王娜. 基于多标签的军事领域命名实体识别[J]. 计算机科学, 2019, 46(S2):9-12.
- [23] 吴雄辉. 政务公文命名实体识别方法研究及应用[D]. 北京:北京化工大学, 2019.
- [24] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019:4171-4186.
- [25] 胡甜甜. 面向政务领域知识抽取的研究与应用[D]. 贵阳:贵州大学, 2022.
- [26] 赵建东,高光来,飞龙. 蒙古语语音合成语料库标注规则的设计[J]. 内蒙古大学学报(自然科学版), 2013, 44

(3):324-328.

[27] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[28] 岳琪, 李想. 基于BERT和双向RNN的中文林业知识图谱构建研究[J]. 内蒙古大学学报(自然科学版), 2021, 52(2):176-184.

(责任编辑 黄宝琦)

Research on Mongolian Domain Term Recognition Based on Bi-LSTM and CRF

Tana, Songyun

(College of Computer Science and Technology, Inner Mongolia Normal University,
Hohhot 010022, China)

Abstract: To address challenges such as morphological complexity and scarce training data in Mongolian terminology recognition, the Bi-LSTM+CRF sequence labeling model is proposed, which integrates bidirectional long short-term memory (Bi-LSTM) and conditional random fields (CRF). Focusing on three terminology categories including policy terms, organizational names, and title names, a manually annotated Mongolian government text dataset is constructed and the BMEO labeling scheme for sequence annotation is employed. Bi-LSTM is utilized by the model to extract contextual semantic features, and CRF is integrated to optimize label transfer constraints. The approach effectively addresses challenges such as Mongolian's lack of case distinction, high frequency of compound words, and rich morphological variations. Experimental results show that the model achieves an overall F_1 score of 93.17% on the test set, outperforming standalone CRF or Bi-LSTM models. The approach significantly enhances Mongolian terminology recognition accuracy, offering a viable pathway for natural language processing research in resource-scarce languages.

Key words: natural language processing; Mongolian language; domain term; term recognition