专题 2023 多语种智能信息处理数据集

TibNER：藏文命名实体识别数据集

周毛克[1,2]，俄见才让[2,3]，道吉才旦[4]，戚肖克[5*]，赵小兵[2,3*]

1. 中央民族大学中国少数民族语言文学学院，北京 100081
2. 国家语言资源监测与研究民族语言中心，北京 100081
3. 中央民族大学信息工程学院，北京 100081
4. 中国民族语文翻译中心（局），北京 100080
5. 中国政法大学法治信息管理学院，北京 102249

摘要： 结构化的语言资源是自然语言处理的重要基础。目前，由于缺乏公开的大规模数据集，藏文命名实体识别研究进展缓慢，成果积累较少。基于此，本文利用实体词典半自动地构建并公开了藏文命名实体识别数据集（TibNER）。为保证数据集质量，对自动标注结果进行了人工校审。TibNER 包含 20096 个句子，平均句长为 44.2069 个音节，标注的实体类型包括人名、地名、组织机构名，三类实体总数达 43678。为了验证数据集的有效性，本文在三个主流的序列标注模型上进行对比测试，最优模型的 F1 值达到 80.60%。经研究，本数据为低资源语言提供了数据构建经验，同时为藏文命名实体识别等任务提供了一定的数据基础。

关键词： 藏语；命名实体识别；实体词典；数据集

数据库（集）基本信息简介

数据库（集）名称	TibNER：藏文命名实体识别数据集
数据通信作者	戚肖克（qixiaoke@cupl.edu.cn）；赵小兵（nmzxb_cn@163.com）
数据作者	周毛克、俄见才让、道吉才旦、戚肖克、赵小兵
数据时间范围	2010–2023年
地理区域	中国
数据量	17.2 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.j00001.01089
基金项目	国家社科基金重大项目（22&ZD035）
数据库（集）组成	TibNER数据集共包括3个数据文件，其中（1）trainTibNER.json是训练集，数据量为16078句/13.73MB；（2）devTibNER.json是验证集，数据量为2009句/1.71MB；（3）testTibNER.json是测试集，数据量为2009句/1.76MB。

引　言

实体（Entity）是人类理解文本和处理文本的基本单元之一。根据 ACE（Automatic Content Extraction）评测界定，实体概念在文本中的指称项有三种形

ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2024.0041.zh

文献 DOI：
10.11922/11-6035.csd.2024.0041.zh
数据 DOI：
10.57760/sciencedb.j00001.01089

文献分类：信息科学

收稿日期：2024-02-20
开放同评：2024-05-21
录用日期：2024-11-19
发表日期：2024-12-20

* 论文通信作者
戚肖克：qixiaoke@cupl.edu.cn
赵小兵：nmzxb_cn@163.com

式，即名词性指称、命名性指称和代词性指称[1]。例如图 1 句子中，实体“切阳什姐”的指称项有三个，其中“中国田径运动员”是名词性指称，“切阳什姐”是命名性指称，“她”是代词性指称。命名实体识别（Named Entity Recognition，NER）任务的主要目标就是识别出自然语言文本中实体的命名性指称，即命名实体。该任务最早是在第六届信息理解会议（Message Understanding Conference，MUC-6）上作为实体关系分类的子任务提出。作为自然语言处理（Natural Language Processing，NLP）的关键基础任务之一，NER 在信息抽取、信息检索、知识图谱、问答系统、机器翻译等下游任务中发挥着重要的作用[2]。

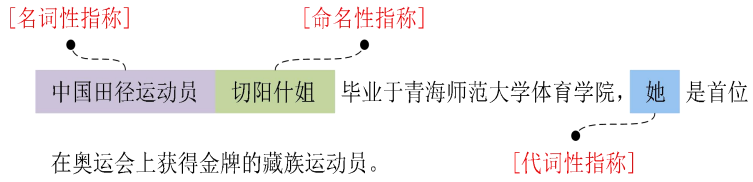

图 1　实体概念的指称项

Figure 1　Referents of entity concepts

藏文命名实体识别（Tibetan Named Entity Recognition，TNER）是藏语 NLP 领域的基础任务。根据文献统计，TNER 研究开始于 2010 年。在近十年的发展中，其研究方法不断更迭，对结构化的语言数据资源的需求逐渐增加。然而，藏语作为低资源语言（Low Resource Language），缺乏公开的大规模、高质量数据集[3]。目前已有的研究成果基本上是基于研究者自建的数据集，如 LIU[4]等构建了 7938 KB 规模的藏文数字识别数据集，在藏文数字识别任务上得到了较好的性能；华却才让[5]、洛桑嘎登[6]等分别构建了规模为 1.6 万句和 5 万句的藏文通用实体识别数据集；加羊吉[7]、珠杰[8]、刘飞飞[9]等分别构建了 5.6 MB、2698 句和 5.3 MB（18789 句）藏文人名识别数据集，并利用 CRF 模型建立藏文人名实体识别模型；环科尤等[10]构建了 10 万句面向格萨尔王史诗的领域命名实体识别数据集。值得强调的是，头旦才让等[11]构建并公开了藏文地名词典及地名实体识别数据集（http://github.com/toudancairang/Tibetan-Computational-linguastics）；邓宇扬等[12]将构建的包含 321 篇文档的藏文节日实体识别数据集（http://github.com/AlieZVzz/CINO-Festival-NER）公开在了 GitHub 平台上供研究者使用。然而，这些公开的数据集规模较小，实体类型单一，不足以训练高性能的 TNER 模型。

针对以上问题，本文提出了一种基于实体词典的 TNER 数据集半自动构建方法。经实体词典自动匹配和人工校审，最终得到规模为 20096 句面向藏文通用领域的命名实体识别数据集 TibNER。本数据集标注了人名（Person，PER）、地名（Location，LOC）、组织机构名（Organization，ORG）三类通用命名实体，为 TNER 研究提供了数据支撑。

1　数据采集和处理方法

1.1　语料采集与预处理

本文的生语料（Raw Corpus）来源于 10 个主流的藏文新闻网站，如人民网（藏文版）、中国西藏新闻网、新华网（藏文版）等，具体如表 1 所示。这些新闻内容涵盖了国内外政治、经济、社会、文化、科技等各个领域，其实体具有“类多量大”的特点。

表 1　生语料来源网站

Table 1　Source websites for raw corpus

序号	网站	URL
1	人民网（藏文版）	http://tibet.people.com.cn/
2	中国西藏网	http://tb.tibet.cn/
3	中国西藏新闻网	http://tb.chinatibetnews.com/
4	新华网（藏文版）	http://xizang.news.cn/
5	青海湖网	http://www.amdotibet.cn/
6	香格里拉藏文网	http://shanggri-latibet.cn/
7	中国藏文网通	https://ti.tibet3.com/
8	中国藏语广播网	http://www.tibetcnr.com/
9	青海省人民政府门户网	https://www.qhtibetan.com/
10	青海藏语网络广播电视台	http://www.qhtb.cn/

为了尽可能地降低语言资源获取与构建成本，本文首先利用网络爬虫技术采集了时间跨度在 2010－2023 年之间的 6189 个藏文新闻文本，并进行数据清洗、去重；然后借助藏文单垂符“།”、双垂符“།།”、终结词（གོང་དུ་རྟོགས་ཤོག་རེ་ལགས།）、离合词（གམ་ངམ་དམ་ནམ་བམ་མམ་འམ་རམ་ལམ་སམ།）以及藏文系动词、存在动词等语尾助词（“ཡིན་”“རེད་”“ཡོད་”“འདུག”“ཐུང་”“སོང་”等）对藏文文本进行分句处理；最后筛选抽取出句子长度介于 2–200 个音节的 56241 句藏文生语料，存储并等待进一步标注。

1.2　藏文命名实体词典

基于实体词典的 NER 数据集构建方法依赖大规模命名实体词典。本文在前期构建的人名、地名、组织机构名 3 类藏文命名实体词典（Tibetan Named Entity Dictionary，TibNED）基础上，继续扩充实体规模。扩充的命名实体来源于中国西藏藏语言文字网的“西藏行政地名词典”“旅游景点词典”（http://cn.zyw.xizang.gov.cn/bmfw/hzdzgjs/）和头旦才让等公开的藏文地名词典（https://github.com/toudancairang/Tibetan-Computational-linguistics）。经过扩充、去重，最终三类实体词典总规模达到 7.2 万余条（https://github.com/maomaomuc/TibetanNamedEntityDictionary），具体的实体规模、实体长度及示例如表 2 所示。

表 2　藏文命名实体词典

Table 2　Tibetan named entity dictionary

TibNED	实体规模/个	词长范围/音节	示例
PERED	28255	1–22	ནམ་མཁའ་ལེགས་པའི་ཐོ་གོས་རྒྱལ་མཚན་དཔལ་བཟང་པོ་ [南喀勒白洛追坚赞贝桑波]
LOCED	38044	1–35	ཡུན་ནན་ཞེང་ཆེན་ཁད་ཆེན་བོད་རིགས་རང་སྐྱོང་ཁུལ་ཞམས་ཀྱི་ཉི་ཟླ་རྫོང་ལོ་སྐྱིད་ཤང་ [云南省迪庆藏族自治州香格里拉县洛吉乡]
ORGED	5980	1–23	སྤྱི་ཚོགས་ཚན་རིག་ལས་ཀྱིའི་སྡ་ཁྱེད་སོོག་ཐོབ་ཁང་ཞེས་བཅེར་ཁག་འབེལ་ལས་ཚུ་ཡོན་རྒྱན་ཁང་ [社会科学工作人员业务职称评审委员会]
总计	72279	1–35	–

1.3 TibNER 构建方法

在以句为单位的藏文生语料和 TibNED 基础上，本文提出了一种基于实体词典匹配的命名实体标注算法，根据生语料的数据量分 6 轮进行自动标注和人工校审，具体的构建流程如图 2 所示：

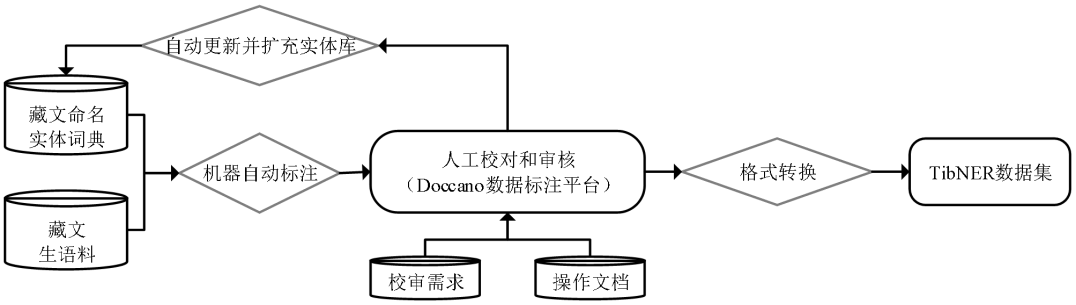

图 2　TibNER 构建流程

Figure 2　The flowchart for the TibNER dataset construction

构建算法如下所示：

算法：基于实体词典匹配的 TibNER 数据集构建算法

输入：藏文句子 TibSent，藏文命名实体词典 TibNED；

输出：标注后的藏文句子 OutTibSent；

初始化：OutTibSent = “”；

步骤 1：TibWordList ← 利用“卓玛拉藏文词法分析软件[①]”对输入的藏文句子 TibSent 进行分词处理；

步骤 2：实体词典匹配算法

```
wordList = []
for word in TibWordsList:
    if word in PERED:
        wordList.append(word + “/PER”)
    elif word in LOCED:
        wordList.append(word + “/LOC”)
    elif word in ORGED:
        wordList.append(word + “/ORG”)
    else:
        wordList.append(word + “/O ”)
wordlist.append(“\n”)
OutTibSent = “”.join(wordlist)
Return OutTibSent
```

步骤 3：删除无命名实体的句子并将包含实体的句子转换成 JSON 格式上传至 Doccano 数据标注平台进行人工校审；

步骤 4：从人工校审结果中抽取实体、去重并更新 TibNED；

步骤 5：返回步骤 1 进行新一轮标注（循环直至结束）；

步骤 6：将经过“标校审”的数据下载并存入 TibNER 数据库中，用于模型训练、验证和测试。

[①] 本文使用的藏文分词软件为中国社会科学院民族学与人类学研究所民族语言文化行为实验研究室构建的“卓玛拉藏文词法分析软件”，该系统在第二届少数民族语言分词评测中获得二等奖，其藏文分词 F1 值达到 94.71%。

经过 6 轮标注、校对和审核工作，最终构建了规模为 20096 句，平均句长为 44.2069 个音节的 TNER 数据集 TibNER，详情如表 3 所示。其中，数据的校对和审核工作是在 Doccano 数据标注平台（http://127.0.0.1:8000/projects/1/）进行，示例如图 3 所示。此外，在构建 TibNER 数据集过程中，每轮“标校审”结束后都会实时更新 3 类 TibNED，更新后的实体词典规模及新增实体情况如表 4 所示。

表 3　TibNER 数据集

Table 3　TibNER dataset

数据集	句对	句长 (avg/max/min)	音节/万	实体类别及规模			实体总数
				人名	地名	组织机构名	
TibNER	20096	44.2069/152/2	88.1075	10610	21034	12034	43678

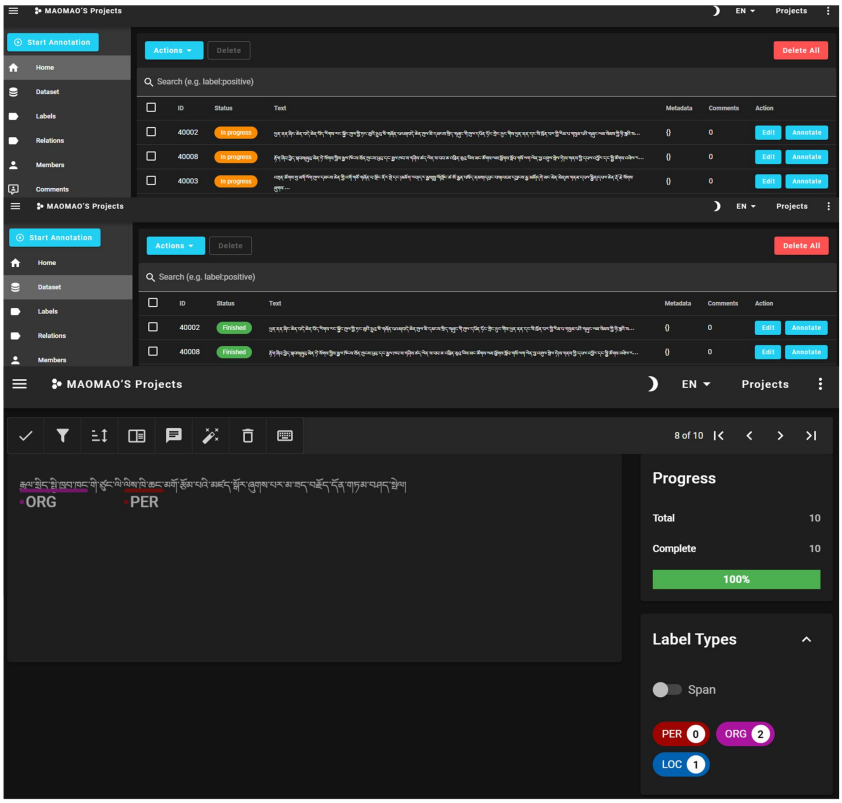

图 3　Doccano 数据校对审核平台

Figure 3　Doccano Data Proofreading and Auditing Platform

表 4　更新后的藏文命名实体词典

Table 4　Updated Tibetan named entities dictionary

TibNED	实体规模	更新后的实体规模	新增量	词长范围
PERED	28255	33467	5212	1–22
LOCED	38044	41979	3935	1–35
ORGED	5980	10171	4191	1–23
总计	72279	**85617**	**13338**	1–35

2 数据样本描述

在 TibNER 数据集构建过程中，本文利用 TibNED 自动匹配标注原始生语料，标注结果如图 4 所示。同时借助 Doccano 数据标注平台对机器标注的数据进行人工校对和审核，校审后导出的数据为 JSON 格式，如图 5 所示。除了将数据保存为 JSON 格式之外，本文还将 JSON 格式数据自动转换成了“BIO”序列标注格式，满足不同模型的数据格式需求，样本示例如图 6 所示。

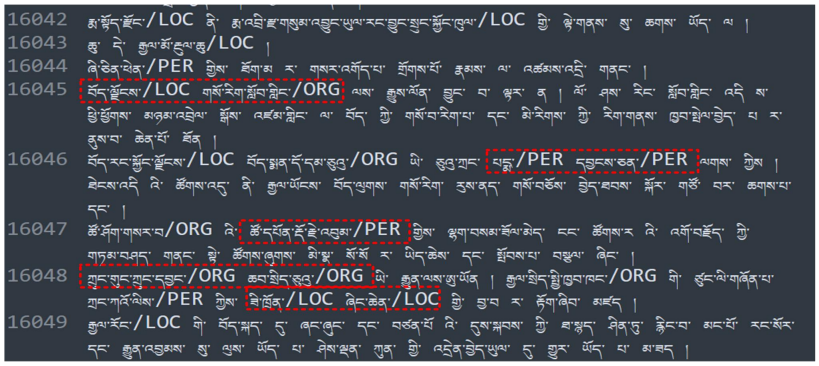

图 4　自动标注后的数据集示例

Figure 4　Dataset example after automatic labelling

图 5　JSON 格式的数据集示例

Figure 5　Dataset example in JSON format

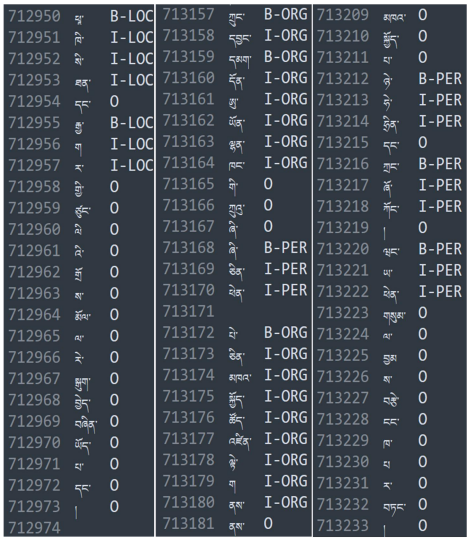

图 6　BIO 格式的数据集示例

Figure 6　Dataset example in BIO format

3 数据质量控制和评估

为了验证数据集的有效性，本文以自建的 TibNER 数据集为基础，在序列标注模型 CRF、BiLSTM、BiLSTM-CRF[13]上进行对比实验。按照 8:1:1 的比例将 TibNER 数据集随机划分为训练集、验证集和测试集，数据详情如表 5 所示。同时采用准确率、召回率和 F1 值作为实验的评价指标。此外，本文的模型是基于 Python3.9.13 和 pytorch1.13.0 框架，代码均在 GPU 服务器上运行，最终的实验结果如表 6 所示。

表 5　数据集规模

Table 5　Dataset size

TibNER	句对	音节/万	句长 (ave/max/min)	实体类型及规模			实体总数
				人名	地名	组织机构名	
训练集	16078	70.2408	43.6875/152/2	8394	16767	9698	34859
验证集	2009	8.7478	43.5430/125/3	1094	2121	1136	4351
测试集	2009	9.1189	45.3902/152/3	1122	2146	1200	4468
总计	20096	88.1075	44.2069/152/2	10610	21034	12034	43678

表 6　实验结果

Table 6　Experimental results

TNER Model	P(%)	R(%)	F1(%)
CRF	80.25	78.17	79.20
BiLSTM	77.42	78.85	78.13
BiLSTM-CRF	79.07	82.19	80.60

根据表 6 中的实验结果可知，本文构建的 TibNER 数据集在 TNER 任务上取得了较好的性能。整体上，BiLSTM-CRF 模型性能最佳，F1 值达到 80.60%，同时该模型泛化能力较强，取得了较高的召回率，值为 82.19%，比 BiLSTM、CRF 模型分别高出 3.34 和 4.02 个百分点。但是，在三个模型中，CRF 模型的准确率最高，值为 80.25%。

除了给出总体的评价，本文还将不同模型在单个实体上的效果也做了对比实验，实验结果如表 7 所示。经实验，三类实体中，地名实体识别效果最佳，人名次之，组织机构名效果相对较低。这种情况的出现与 TibNER 数据集的实体分布、实体特点等有一定的关联。图 7 给出了数据集实体分布情况。从图中可以看出，地名实体数约为人名或组织机构名的 2 倍，在训练过程中，由于地名样本丰富，模型能够充分学习地名特征，这也是地名效果最佳的原因之一。此外，人名识别效果优于组织机构名是因为人名实体长度较短且固定，人名一般以单音节、双音节、三音节及四音节为主，大于 4 个音节的人名较少，而组织机构名长度不固定，最长的能够达到 20 个音节并且容易与地名实体混淆。

综上，在 TibNER 数据集上，最优模型的 F1 值能够达到 80.60%，但实验结果仍然存在一些问题，如细粒度分词导致的实体识别错误问题、实体与藏文功能性虚词黏写现象导致的错误以及多层嵌套实体识别结果不一致问题等。经过分析，这类错误一方面是由于模型泛化能力和算力导致，另一方面与数据集构建过程中分词器的分词粒度及数据集规模、质量有直接的关联，具体示例如图 8 所示。

表 7　单个实体上的实验结果

Table 7　Experimental results on single entities

TNER Models	实体类型								
	人名（PER）			地名（LOC）			组织机构名（ORG）		
	P	R	F1	P	R	F1	P	R	F1
CRF	82.83	81.57	82.19	82.95	81.36	82.14	74.97	71.59	73.24
BiLSTM	79.33	81.23	80.26	79.47	82.61	81	73.47	72.73	73.09
BiLSTM-CRF	80.11	86.09	82.99	82.67	84.13	83.39	74.43	76.35	75.37

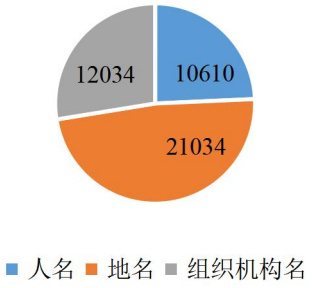

图 7　TibNER 数据集实体分布情况

Figure 7　Distribution of entities in the TibNER dataset

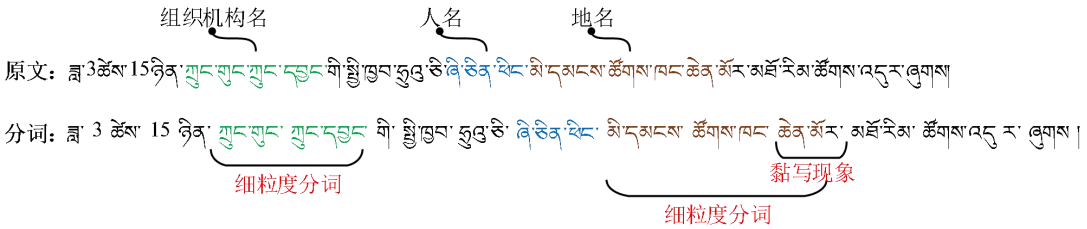

图 8　错误案例

Figure 8　Error cases

总之，在 TibNER 数据集构建及模型对比过程中发现，除了算法、算力之外，数据集的规模、质量等因素也是直接影响 TNER 结果的重要因素，在接下来的研究中，除了更好地把握算法和算力之外，还需要在构建大规模高质量数据集上加大力度。

4　数据价值

藏语低资源属性一定程度上限制了藏语 NLP 的发展进程。针对目前藏文缺少公开的结构化命名实体识别数据集的问题，本文半自动地构建了 20096 句规模的 TibNER 数据集。经过实验，以 TibNER 为基础的 TNER 模型的 F1 值最高达到 80.60%。实验结果证明，基于 TibNER 数据集训练的 TNER 模型具有较好的性能。因此，本数据集的构建不仅为低资源语言提供了数据集构建经验并且为 TNER、藏文信息抽取、知识图谱构建等藏语自然语言处理工作提供了一定的数据基础，具有一定的现实意义与应用价值。

致　谢

感谢中国社会科学院民族学与人类学研究所民族语言文化行为实验研究室对本数据集分词部分提供藏文分词软件（卓玛拉藏文词法分析软件）。

数据作者分工职责

周毛克（1993—），女，甘肃省甘南藏族自治州夏河县人，博士研究生，研究方向为自然语言处理。主要承担工作：数据获取与预处理、数据校对与审核、论文撰写。

俄见才让（1994—），男，青海省海南藏族自治州共和县人，硕士研究生，研究方向为自然语言处理。主要承担工作：数据校对与审核。

道吉才旦（1993—），男，甘肃省甘南藏族自治州合作市人，硕士，研究方向为藏学（藏汉翻译）。主要承担工作：数据校对与审核。

戚肖克（1985—），女，山东省菏泽市人，博士，副教授，研究方向为语音信号处理、自然语言处理。主要承担工作：论文写作指导。

赵小兵（1967—），女，内蒙古自治区呼和浩特市人，博士，教授，研究方向为自然语言处理、语言信息安全。主要承担工作：研究思路设计。

参考文献

[1] 赵军, 刘康, 何世柱, 等. 知识图谱[M]. 北京: 高等教育出版社, 2018. [ZHAO J, LIU K, HE S Z, et al. Knowledge graph[M]. Beijing: Higher Education Press, 2018.]

[2] YADAV V , BETHARD S . A Survey on Recent Advances in Named Entity Recognition from Deep Learning models[C]//.In Proceedings of the 27th International Conference on Computational Linguistics(COLING2018), Santa Fe, New Mexico, USA, 2018:2145-2158.

[3] LONG C J, HILL N W. Recent developments in Tibetan NLP[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2021, 20(2): 1–3. DOI: 10.1145/3453692.

[4] LIU H D, ZHAO W, NUO M H, et al. Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation[C]//. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING2010),2010:719-724.

[5] 华却才让, 姜文斌, 赵海兴, 等. 基于感知机模型藏文命名实体识别[J]. 计算机工程与应用, 2014, 50(15): 172–176. DOI: 10.3778/j.issn.1002-8331.1308-0196. [HUA Q, JIANG W B, ZHAO H X, et al. Tibetan Name entity recognition with perceptron model[J]. Computer Engineering and Applications, 2014, 50(15): 172–176. DOI: 10.3778/j.issn.1002-8331.1308-0196.]

[6] 洛桑嘎登, 群诺, 索南尖措, 等. 融合音节部件特征的藏文命名实体识别方法[J]. 厦门大学学报(自然科学版), 2022, 61(4): 624–629. DOI: 10.6043/j.issn.0438-0479.202110013. [LUO S, QUN N, SUO N, et al. Fusion of syllable component features for Tibetan named entity recognition[J]. Journal of Xiamen University (Natural Science), 2022, 61(4): 624–629. DOI: 10.6043/j.issn.0438-0479.202110013.]

[7] 加羊吉, 李亚超, 宗成庆, 等. 最大熵和条件随机场模型相融合的藏文人名识别[J]. 中文信息学

报, 2014, 28(1): 107 – 112. DOI: 10.3969/j.issn.1003-0077.2014.01.015. [JIA Y J, LI Y C, ZONG C Q, et al. A hybrid approach to Tibetan person Name identification by maximum entropy model and conditional random fields[J]. Journal of Chinese Information Processing, 2014, 28(1): 107 – 112. DOI: 10.3969/j.issn.1003-0077.2014.01.015.]

[8] 珠杰, 李天瑞, 刘胜久. 基于条件随机场的藏文人名识别技术研究[J]. 南京大学学报(自然科学), 2016, 52(2): 289 – 299. DOI: 10.13232/j.cnki.jnju.2016.02.010. [ZHUJIE, LI T R, LIU S J. Research on Tibetan Name recognition technology under CRF[J]. Journal of Nanjing University (Natural Sciences), 2016, 52(2): 289 – 299. DOI: 10.13232/j.cnki.jnju.2016.02.010.]

[9] 刘飞飞, 王志娟. 基于层次特征的藏文人名识别研究[J]. 计算机应用研究, 2018, 35(9): 2583 – 2587, 2596. DOI: 10.3969/j.issn.1001-3695.2018.09.005. [LIU F F, WANG Z J. Research on recognition of Tibetan names based on hierarchical features[J]. Application Research of Computers, 2018, 35(9): 2583 – 2587, 2596. DOI: 10.3969/j.issn.1001-3695.2018.09.005.]

[10] 环科尤, 华却才让, 才让当知, 等. 基于深度学习的格萨尔史诗命名实体识别研究[J]. 中文信息学报, 2022, 36(8): 46 – 53. DOI: 10.3969/j.issn.1003-0077.2022.08.006. [HUAN K Y, HUA Q, CAI R, et al. Reseorch of gesar epic named entity recognition based on deep learning[J]. Journal of Chinese Information Processing, 2022, 36(8): 46 – 53. DOI: 10.3969/j.issn.1003-0077.2022.08.006.]

[11] 头旦才让, 仁青东主, 尼玛扎西. 基于 CRF 的藏文地名识别技术研究[J]. 计算机工程与应用, 2019, 55(18): 111 – 115. DOI: 10.3778/j.issn.1002-8331.1903-0232. [TOUDAN C R, RENQING D Z, NIMA Z X. Research on Tibetan location Name recognition technology under CRF[J]. Computer Engineering and Applications, 2019, 55(18): 111 – 115. DOI: 10.3778/j.issn.1002-8331.1903-0232.

[12] 邓宇扬, 吴丹. 面向藏族传统节日的汉藏双语命名实体识别研究[J]. 数据分析与知识发现, 2023, 7(7): 125 – 135. DOI: 10.11925/infotech.2096-3467.2022.0698. [DENG Y Y, WU D. Chinese-Tibetan bilingual named entity recognition for traditional Tibetan festivals[J]. Data Analysis and Knowledge Discovery, 2023, 7(7): 125 – 135. DOI: 10.11925/infotech.2096-3467.2022.0698.]

[13] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. 2015: 1508.01991.http://arxiv.org/abs/1508.01991v1

论文引用格式

周毛克, 俄见才让, 道吉才旦, 等. TibNER：藏文命名实体识别数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-20). DOI: 10.11922/11-6035.csd.2024.0041.zh.

数据引用格式

周毛克, 俄见才让, 道吉才旦, 等. TibNER：藏文命名实体识别数据集[DS/OL]. V2. Science Data Bank, 2024. (2024-12-19). DOI: 10.57760/sciencedb.j00001.01089.

A dataset of Tibetan named entity recognition—TibNER

ZHOU Maoke[1,2], EJIAN Cairang[2,3], DAOJI Caidan[4], Qi Xiaoke[5*], ZHAO Xiaobing[2,3*]

1. School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, P.R. China

2. National Language Resource Monitoring & Research Center of Minority Languages, Beijing 100081, P.R. China

3. School of Information Engineering, Minzu university of China, Beijing 100081, P.R. China

4. China Ethnic Languages Translation Centre, Beijing 100081, P.R. China

5.China University of Political Science and Law, Beijing 102249, P. R. China

*Email:qixiaoke@cupl.edu.cn, nmzxb_cn@163.com

Abstract: Structured linguistic resources are an important foundation for natural language processing. Due to the lack of publicly available large-scale datasets, few researches focus on Tibetan Named Entity Recognition, with limited achievements. Therefore, in this paper we semi-automatically construct a dataset of Tibetan Named Entity Recognition named “TibNER” based on an entity dictionary. To ensure the quality of the dataset, we manually proofread the automatic annotation results. TibNER contains 20,096 sentences with an average sentence length of 44.2069 syllables, and the labelled entities include speaker names, place names and organization names, with a total of 43,678 entities across these categories. In order to validate the dataset, we tested it on three mainstream sequence annotation models, and the best model had an F1 value of 80.60%. According to studies, this dataset can not only provide data construction experience for low-resource languages, but also serves as a solid data foundation for researches like Tibetan named entity recognition.

Keywords: Tibetan; named entity recognition; entity dictionary; dataset

Dataset Profile

Title	A dataset of Tibetan named entity recognition—TibNER
Data corresponding author	Qi Xiaoke (qixiaoke@cupl.edu.cn), ZHAO Xiaobing (nmzxb_cn@163.com)
Data authors	ZHOU Maoke, EJIAN Cairang, DAOJI Caidan, Qi Xiaoke, ZHAO Xiaobing
Time range	2010–2023
Geographical scope	China
Data volume	17.2 MB
Data format	*.json
Data service system	<https://doi.org/10.57760/sciencedb.j00001.01089>
Source of funding	National Social Science Foundation of China (22&ZD035)
Dataset composition	TibNER Dataset consists of 3 subsets in total. (1) trainTibNER.json is the training composed of 16,078 sentences with a data volume of 13.73MB; (2) devTibNER.json is the validation set composed of 2009 sentences, with a data volume of 1.71MB; (3) testTibNER.json is the test set composed of 2,009 sentences with a data volume of 1.76MB.