基于深度学习的格萨尔史诗命名实体识别研究

环科尤[1,2,3*] 华却才让[1,2,3] 才让当知[1,2,3] 多杰才让[1,2,3]

（1.青海师范大学 计算机学院，青海 西宁 810016；
2.藏语智能信息处理及应用国家重点实验室，青海 西宁 810008；
3.青海省藏文信息处理与机器翻译重点实验室，青海 西宁 810008）

摘要：为深入研究藏文命名实体的基本结构，分类方法以及自动识别技术，同时为进一步完善藏语词法分析、句法和语义分析，以及机器翻译等信息处理领域的基础性研究工作。该文专门研究了实体类型较为丰富的格萨尔史诗文本，制定了六种格萨尔命名实体类型，提出了藏文音节和深度学习结合的格萨尔史诗命名实体识别方法。在人工标注的 10 万多句命名实体训练集和测试集上，经实验命名实体识别的准确率、召回率和 F 值分别达到 96.87%、97.11% 和 96.99%。满足了研究格萨尔史诗命名实体识别的应用需求，同时为格萨尔知识图谱和藏文命名实体标注规范的制定提供了依据。

关键词：格萨尔；命名实体识别；深度学习

中图分类号：TP391 **文献标识码：**A **文章编号：**

1 引言

命名实体识别是自然语言处理任务中的重要基础性工作之一，其主要目的是识别给定文本中的命名实体。还可用于处理很多下游 NLP 任务，例如句法分析、关系提取、事件提取、问答系统和机器翻译等。而格萨尔史诗经典版本《霍岭》中史诗人物超过了 1000 人、场景或故事地点达 800 多个,生活用具 1000 多种,武器铠甲等 400 多种,甚至战马名称也多达 140 多个,战神等神祇更是多达 400 多个[1]。若对实体如此庞杂的史诗语料做命名实体自动识别处理，将有助于提升下游藏文信息处理领域的质量。

目前,在英文和汉文的命名实体识别方面相关研究者已经做了许多研究，并且研究内容和实验成果也相对很好[2-8]。而藏文命名实体识别研究中，于洪志、窦嵘和金明等融合人了词典和基于规则方法进行了藏文命名实体识别的初步尝试，是最早可查的关于此领域的研究成果。后来，加羊吉、华却才让和珠杰等用了基于混合和基于神经网络的方法来研究藏文命名实体。其中，华却才让、加羊吉、刘飞飞和贡保才让等研究了三大类的藏文命名实体。其他研究者研究了一类实体的识别性能。对比以上论文及实验结果表明，目前研究三大类藏文命名实体识别中最高的综合 F 值为 89.09%[9]，以及研究单一类藏文命名实体识别中最高的 F 值分别是藏文机构名为 91.09%[10]、藏文人名为 88.30%[11]、藏文地名为 88.45%[12]。

除上述命名实体识别研究的不同方法,各实验数据的内容及规模也有所不同，尤其是藏文命名实体的类型与数量部分，藏文命名实体的分类仅限于人名、地名和机构名，识别类型相对较少，未见到针对特定领域命名实体自动识别的研究文献，为此藏文命名实体识别存在进一步研究和提升的空间，故为进一步丰富藏文命名实体的研究领域，补充和构建不同领域的语料数据资源库，本研究以具有丰富藏文命名实体类型的格萨尔史诗，主要以格萨尔史诗经典版本《北方降魔》、《霍岭大战》、《闷岭大战》和《姜岭大战》等著名的四大降魔史[13]的文献资源为基础，提出了以藏文音节（Tibetan Syllable,以下简称“TS”）为基本单元的 TS-BILSTM-CRF 的格萨尔史诗命名实体识别（简称 GesarNER）方法，将格萨尔文献中的命名实体归纳总结，并分为六种类型，以半自动方式标注了较大规模的格萨尔命名实体语料库，经实验取得了良好的结果。

收稿日期：xxxx-xx-xx **录用日期：**xxxx-xx-xx

基金项目：互联网+藏语信息处理平台建设项目(2017-GX-146)；藏语智能信息处理及应用国家重点实验室项目(2020-ZJ-Y05)；面向农牧区的藏语智能语音交互关键技术研究(2019-SF-129)。

* **通信作者：**huaquer23@qq.com

2 格萨尔命名实体类型及特征

格萨尔史诗的命名实体非常丰富，该史诗中除了包含传统的人名、故事地名、组织机构名等的命名方式外，还包括其独居特色的神祇、神兽坐骑、武器铠甲等实体，每种实体赋予了独特的文化背景。本文依据格萨尔史诗的命名实体的特征，制定了六种格萨尔命名实体类型，具体内容如下：

2.1 人名实体的特征

人名实体以人类为区分个体，给每个个体给定的特定名称符号，即每个人都有的一种代号[14]。本文研究的格萨尔史诗中人名实体主要包括人名和神祇名，其结构特征以六、四音节为主。

在格萨尔史诗和格萨尔大辞典及文献中职衔名、地名等对人名的修饰归为人名实体的组成部分，为此本规范以人名关联的职衔名和部落名归为人名实体。根据以上格萨尔人名实体的结构一般可分为十五种方式命名。（1）“人名”命名的，例如，“དོན་གྲུབ་དཀར་པོ།”（敦珠噶布），（2）“人名+赞颂名”命名的，例如，“ནེའུ་ཆུང་ལུ་གྱའི་ཚར་ཡག”（莱琼·鲁姑查娅），（3）“人名+职衔名”命名的，例如，“གེ་སར་རྒྱལ་པོ།”（格萨尔王），（4）“人名+赞颂名+职衔名”命名的，例如，“གེ་སར་ནོར་བུ་དཔའ་འདུལ་རྒྱལ་པོ།”（格萨尔宝珠制敌王），（5）“姓氏+部落名+人名”命名的，例如，“སྡོང་འབྲུམ་པའི་རྒྱ་ཚ་ཞལ་དཀར།”（董·奔巴·嘉察协噶），（6）“部落名+人名”命名的，例如，“གླིང་ཚ་ཞང་འདོན་མ་ཐང་ཁ།”（岭·察香丹玛绛查），（7）“部落名+职衔名+人名”命名的，例如，“སྤུག་རོང་དོབོན་པོ་ཁོ་ཐུང་”（达绒长官晁通），（8）“部落名+职衔名+人名+职衔名”命名的，例如，“གླིང་རྩེ་གེ་སར་རྒྱལ་པོ།”（岭国大王格萨尔王），（9）“部落名+人名+职衔名”命名的，例如，“གླིང་གེ་སར་དཔའ་འདུལ་རྒྱལ་པོ།”（岭国格萨尔宝珠王），（10）“职衔名+人名”命名的，例如，“ཤན་པ་མེའུ་རྩེ།”（辛巴·梅乳泽），（11）“职衔名+部落名+人名”命名的，例如，“སྨྱི་དཔོན་རོང་ཚ་ཁ་ཤན།”（总管王绒察查根），（12）“赞颂名”命名的，例如，“མེང་ཆེན།”（桑庆），（13）“赞颂名+人名”命名的，例如，“སྐ་སྲས་ཇོ་རྒྱ།”（噶萨觉如），（14）“赞颂名+人名+职衔名”命名的，例如，“འཛམ་གླིང་གེ་སར་རྒྱལ་པོ།”（世界格萨尔大王），（15）“赞颂名+人名+赞颂名+职衔名”命名的，例如，“འཛམ་གླིང་མེང་ཆེན་སྐྱེས་བུ་ཆེན་པོ་གེ་སར་ནོར་བུ་དཔའ་འདུལ།”（世界雄狮·格萨尔宝珠制敌王）。

除此之外，格萨尔史诗中的大多数人名中有一人多名的现象。如格萨尔王有 38 种别名[15]。格萨尔人名的长度不固定，最长由 14 个音节组成。如“འཛམ་གླིང་མེང་ཆེན་སྐྱེས་བུ་ཆེན་པོ་གེ་སར་ནོར་བུ་དཔའ་འདུལ།”（译为：世界雄狮·格萨尔宝珠制敌王。）。本文中以 PER（Person name）表示人名实体。

2.2 地名实体的特征

地名实体是指人们赋予某一特定空间位置上对自然或人文地理实体的专有名称[16]。本文研究的格萨尔史诗中，地名实体主要以四、六音节为主，包括场景、宫殿和城堡。

在格萨尔史诗中，将山水、地点等的名称合到一起命名为地名，格萨尔地名实体跟一般藏文地名实体一样，地名的内部结构上没有明显的约束，主要用实体外部的上下文信息来确定是否为地名实体。如，“སྒུ་མཁར་མེང་འབྲུག་སྟག་ཅེ།”（狮龙宫殿）、“གངས་རི་ཤེལ་རྫོང་།”（雪山水晶宗）、“འདན་མའི་ནས་རྫོང་།”、（丹玛青稞宗）“མཁར་དུང་དཀར་བསམ་འགྲུབ།”（白螺如意城堡）等，地名的后缀词主要关联以域、小国、城堡(城)、宗、岱等上下文信息。另外，一地多名的现象也颇多。如童门、富饶地、富饶岭国都是岭域的别名。本文中以 LOC（Location name）表示地名实体。

2.3 组织机构名实体的特征

组织机构名一般由三部分组成，即区域名、矢名和格级名，它们分别说明和规定着机构的管理(服务)范围、隶属关系、工作内容以及级别规格等[17]，如中国格萨尔研究中心，中国为区域名，格萨尔为矢名，研究中心为级别规格名。但在格萨尔史诗中，组织机构名实体的长度不固定，未定由三部分组成。主要以单音节、双音节、六音节为主，包括部落名、族群、寺院等。如，“གླིང་།”（岭）、“སྒུ་ལོ།”（嘉洛）、“སྡོད་གླིང་གསེར་པ་ཆེ་རྒྱུད།”（长系）“བར་གླིང་འོམ་བུ་ཆོ་རྒྱུ་”（中系）、“སྨད་གླིང་སྐུ་ཤྲུང་སྡེ་བཞི།”（幼系）等，组织机构名的后缀词主要关联以部落名、人名、其他实体等上下文信息。本文中以 ORG（Organization name）表示人名实体。

2.4 武器铠甲实体的特征

武器铠甲实体是指英雄的兵器，是用于攻击的工具，也因此被用来威慑和防御。在格萨尔史诗中武器铠甲实体主要以双音节、四音节为主，包括战旗及魔法名。如，“རལ་གྲི་ཡ་ཟི་སྐར་ཟན།”（亚斯流

星剑）、“བཀྲ་ཤ་ལན་མེད།”（利钦大刀）、“ཕྱ་མདའ་རྩེ་བརྒྱ།”（尊箭）、“རྟོ་ཝི་ཡང་འཕྲེལ།”（陀罗尼（咒语））等，格萨尔史诗中武器铠甲实体的类型和命名丰富多彩，这类命名实体需要独立地分类到史诗实体类型的分支之一。本文中以 WEA（Weaponry Armor）表示武器铠甲实体。

2.5 神兽坐骑实体的特征

神兽坐骑实体是指英雄的护身物和战马，由于在格萨尔史诗中神兽和坐骑有特殊命名，此实体类型包括各种神兽和坐骑等的名称。格萨尔史诗的坐骑命名实体是整个藏文典籍中最丰富也是唯一的马名资源。在格萨尔史诗中神兽坐骑实体主要以四音节为主，大多数英雄的战马都有与众不同的名字，如“རྨ་རྩོད་འཕེར་བ།”（赤兔马）、“རྒྱ་ཐ་སོག་དཀར།”（加夏素噶）、“གཡེར་མདངས་ཉི་མའི་འཕྲལ་འཁོར།”（金额太阳幻轮马）、“ཧམ་འཇིགས་རྡེད་མོང་སོག་དཀར།”（白肩熊）等，其中也有一马多名的现象。本文中以 HER（Heroic Riding）表示神兽坐骑实体。

2.6 生活用具实体的特征

生活用具实体指的是人类生活的文化（工具）记载到术语的象征，是民俗印迹。如在格萨尔史诗中生活工具有“གཡེར་སྐྱ་ཉི་མ་རངས་ཤར།”（金鞍尼玛让秀）、“གཡེར་རྩོག་ཉི་མའི་འོད་སྣང་།”（金盔）、“བཅུད་འཇོན་རངས་ཐབག”（磨盘石）、“རྨ་ངུའི་རྩིག་རྩེས།”（马蹄石）等。本文中以 LIV（Livelihood Symbol）表示生活用具实体。

格萨尔史诗是多领域研究的知识资源，本文从命名实体识别的角度来探索及研究，并制定了六种类型的格萨尔命名实体。其中，第六种实体类型是生活用具及需要分类但还未分类的其它实体组成，这种类型的实体特征暂时无法详细地归纳，需要进一步研究。

3 数据预处理及实体边界算法

3.1 数据预处理方法

在自然语言处理中，词向量表示也称分布式表示（Distributed Representation）或词嵌入（Word Embedding），其目标是将语言的基本单元用最优化的向量表示，以便计算机能够更好地理解自然语言[18]。藏文词向量表示的语言单元为构件、字丁、字、音节和词等，其中音节作为词向量的语言单元，其预处理效果优于其它语言单元，只要解决藏文紧缩词对音节边界的影响即可。

藏文音节是指以藏文音节点为界限的藏文字符组合，是由藏文字丁构成的最小的语言基本单位[19]。以藏文音节为基本单元的语料处理方法有两种，一种是按音节点和特殊符号来切分音节；另一种是在第一种方法的基础上采用规则和还原法来切分紧缩音节。前者保留了原语料的完整正结构和文法接续特征，对信息抽取和句法分析、语义分析等自然语言处理的下层任务有很好的支撑作用。后者虽然较大地提升了分词的效果，但会破坏原语料的语言结构，出现多余的字词，可以说不符合藏文语法规律。因此，本文的预处理方法是第一种方法，其方法没有改变原语料的语法规律、处理效率优于第二种方法，省时省力。两种方法的结果如表 1 所示。

表 1 藏文音节基本单元的切分结果对比

Table1 Comparison of the results of basic units of Tibetan syllables

原语料	ཐྲིང་གི་སྲིད་ཞིར་མངའ་གསོལ་བའི་རྒྱ་ཚའི་སྲས་ནི་གེ་སར་གྱི་དབོན་པོའམ་ཚ་བོའང་ཡིན་པས་ཀུན་ཀྱང་དགའོ།
中文释义	即位岭国司政者嘉察之子，也是格萨尔王的侄子，故皆大欢喜。
第一种方法	ཐྲིང་ གི་ སྲིད་ ཞིར་ མངའ་ གསོལ་ བའི་ རྒྱ་ ཚའི་ སྲས་ ནི་ གེ་ སར་ གྱི་ དབོན་ པོའམ་ ཚ་ བོའང་ ཡིན་ པས་ ཀུན་ ཀྱང་ དགའོ །
第二种方法（规则法）	ཐྲིང་ གི་ སྲིད་ ཞི ར་ མངའ་ གསོལ་ བ འི་ རྒྱ་ ཚ འི་ སྲས་ ནི་ གེ་ སར་ གྱི་ དབོན་ པོ འམ་ ཚ་ བོ འང་ ཡིན་ པ ས་ ཀུན་ ཀྱང་ དགའ་ བོ །
第二种方法（还原法）	ཐྲིང་ གི་ སྲིད་ ཞི་ ལ་ མངའ་ གསོལ་ བ་ ཡི་ རྒྱ་ ཚ་ ཡི་ སྲས་ ནི་ གེ་ སར་ གྱི་ དབོན་ པོ་ འམ་ ཚ་ བོ་ འང་ ཡིན་ པ་ ཡིས་ ཀུན་ ཀྱང་ དགའ་ བོ་ །

3.2 识别出的实体边界算法

为了解决藏文紧缩词对实体边界的问题，本文在结合两种方法的优点上改进了优化算法：首先，第一种方法作为模型的音节向量化；其次，用音节向量化结合长短期记忆模型和条件随机场

来识别出命名实体，包括实体识别模型的训练；最后，再运用第二种方法和实体后处理算法来解决紧缩音节的切分问题，从而保证了原语料和命名实体的完整性。其算法如下：

算法 1 格萨尔史诗的实体边界算法：

Input : 读入格萨尔史诗 Text
Output: 抽取格萨尔史诗 Named entity
1. entitys 识别 Text 中所有实体
2. outputs []
3. θ <- 实体词缀是紧缩音节
4. ψ <- 含有词缀 ra 的特殊实体集合
5. φ <- 含有词缀 sa 的特殊实体集合
6. **For** entity ∈ entitys **Do**
7. words = entity
8. **If** words[-1][-2:] in θ **Then**
9. outputs.add(entity[:-2])
10. **Else If** words[-1] not in ψ and words[-1][-1]=="ra" **Then**
11. outputs.add(entity[:-1])
12. **Else If** words[-1] not in φ and words[-1][-1]=="sa" **Then**
13. outputs.add(entity[:-1])
14. **Else**
15. outputs.add(entity)
16. **Return** Outputs

4 基于 TS-BILSTM-CRF 的 GesarNER 模型

4.1 长短期记忆模型

长短期记忆（Long short-term memory，LSTM）是一种特殊的 RNN，它能够在更长的序列中有更好的表现。同时能解决长序列训练过程中的梯度消失和梯度爆炸问题。LSTM 单元结构如图 1 所示。

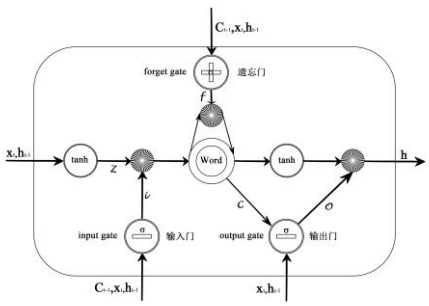

图 1 长短期记忆的单元结构
Figure 1 A Long Short-Term Memory Word

LSTM 的核心主要包括遗忘门、输入门、输出门以及记忆 Word。输入门与遗忘门两者的共同作用就是舍弃无用的信息，把有用的信息传入到下一时刻。对于整个结构的输出，主要是记忆 Word 的输出和输出门的输出相乘所得到的。其结构用公式表达如下。

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \quad (1)$$

$$z_t = tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \quad (2)$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \quad (4)$$

$$0_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right) \quad (5)$$

$$h_t = o_t tanh\left(c_t\right) \quad (6)$$

其中，σ 是 sigmoid 函数。W 是权重矩阵，b 是偏置向量，i_t, f_t, o_t 分别是输入门、遗忘门及输出门和单元向量，z_t 是待增加的内容，c_t 是 t 时刻的更新状态，h_t 则是整个 LSTM 单元 t 时刻的输出。在序列标记中可以在给定的时间内同时获得过去和未来的输入特征，因此我们利用了[20]中提出的双向 LSTM 模型。下图 2 为双向长短期记忆模型结构图。

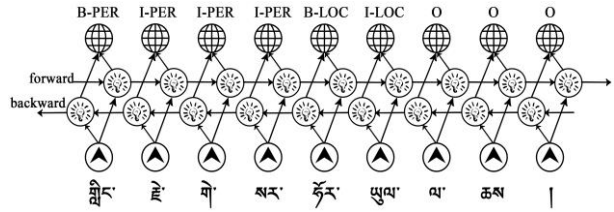

图 2 双向长短期记忆模型
Figure 2 A Bidirectional Long Short Term Memory Model

4.2 TS-BILSTM-CRF 模型

在预测当前标签时，BILSTM 善于处理长距离的上下文信息，但无法处理标签间的依赖信息。CRF 相比其他概率图模型能够利用更加丰富的标签分布信息，能通过邻近标签的关系获得一个最优的预测序列，并弥补 BILSTM 的缺点。本实验将 TS-BILSTM 模型与 CRF 模型相结合，形成一个 TS-BILSTM-CRF 模型如图 3 所示。

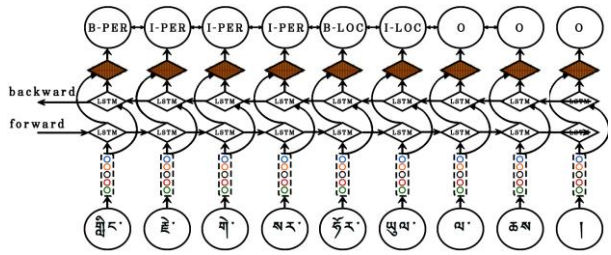

图 3 TS-BILSTM-CRF 模型
Figure 3 A TS-BILSTM-CRF Model

5 实验

5.1 实验数据及其规模

实验数据为格萨尔史诗《北方降魔》、《霍岭大战》、《闷岭大战》和《姜岭大战》等四大降魔史为主的文本语料。数据处理经过了三个步骤，首先，把收集的语料切分为句子级别的文本。其次，把句子级别的文本进行基于音节模式的分词。最后，在基于音节分词的数据中识别有意义的词汇或命名实体。最后人工标注构建了规模达 10 万多句的实体语料库。实验数据中训练集占 80%，剩余作为测试集和开发集，分别对数据中的史诗人物（PER）、史诗地名（LOC）、史诗部落名（ORG）、武器铠甲（WEA）、神兽坐骑（HER）和生活用具（LIV）进行识别，其具体的数据统计如表 2 所示。

表 2 实验数据统计表

Table2 Statistical table of Experimental data

实体类型	训练集	测试集	所有实体
人名（PER）	19734	4919	24653
地名（LOC）	10568	3198	13766
机构名（ORG）	2162	526	2688
武器铠甲（WEA）	989	387	1376
神兽坐骑（HER）	986	312	1298
生活用具（LIV）	2425	556	2981

5.2 实验参数设置及其评价指标

通过多次试验来优化参数，最终各个参数设置如下：字嵌入向量维度设置为 300；优化算法设置为随机梯度下降法的扩展(Adaptive moment estimation，Adam)；模型训练次数设置为 2500；批量处理个数设置为 100；隐藏层的层数设置为 2；隐藏层神经单元个数设置为 256；学习率初始化设置为 0.001；为了防止双向长短期记忆模型过拟合问题，在各模型的输入输出中采用 Dropout，取值为 0.5。图 4 是基本超参数不变只有不同迭代时刻的准确率。

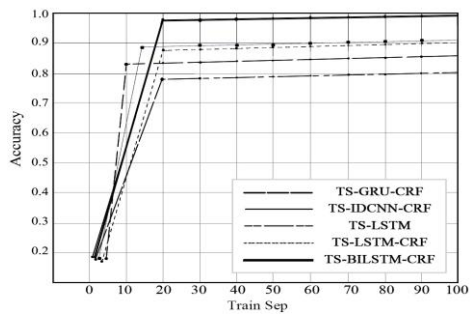

图 4 不同迭代时刻的准确率

Figure 4 Accuracy at different iterations

由图 4 可知，在有限的训练次数（即 2500 次）内 TS-BILSTM-CRF 优于 TS-GRU-CRF、TS-IDCNN-CRF、TS-LSTM 和 TS-LSTM-CRF 的结果，所以，在该实验中选用了双向 LSTM 和 CRF 结合的方法。其他的超参数也是通过这种对比法来选取的。

本实验采用准确率 P（Precision）、召回率 R（Recall）以及 F1(F-Score)值来评判模型的性能[21]。3 个评价指标的计算公式定义如下：

$$P(\text{准确率}) = \frac{\text{正确识别的GESAR实体个数}}{\text{GESAR实体总数}} \times 100\% \quad (7)$$

$$R(\text{召回率}) = \frac{\text{正确识别的GESAR实体个数}}{\text{识别出的GESAR实体总数}} \times 100\% \quad (8)$$

$$F1 = \frac{2 \times \text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \times 100\% \quad (9)$$

5.3 实验结果及其分析

（1）实体识别实验需要一个初始词向量，使用大量文本语料训练词向量效率更高。本实验把藏文音节作为模型的词向量单位，用 Word2vec 语言模型 CBOW，其实验训练 100、200、300 以及 400 维度的向量分别进行对比，实验发现，向量维度过高，实验数据中的噪声容易被捕获，出现过拟合情况；向量维度过低，获取的特征信息不完整，产生欠拟合状况。因此本实验的向量维度为 300，结果如图 5 所示。

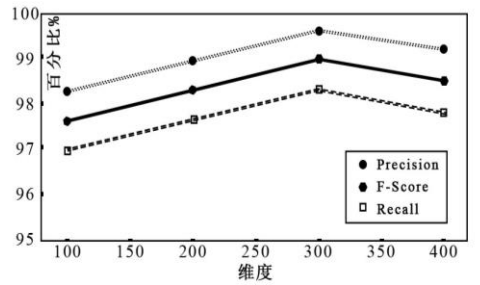

图 5 不同维度词向量实验对比

Figure 5 Word vector experiment comparison

（2）已公开的藏文命名实体识别结果中，研究三类藏文命名实体识别最高 F 值为 89.09%[9]，研究单一类藏文命名实体识别最高 F 值分别是组织机构名为 91.09%[10]、人名为 88.30%[11]、地名为 88.45%[12]。相比已公开的实验结果，基于 TS-BILSTM-CRF 模型的人名、地名、组织机构名、以及三类藏文命名实体识别的 F 值分别提升了 8.61 个百分点(PER)、8.89 个百分点(LOC)、6.47 个百分点(ORG)和 8.18 个百分点(ALL)。武器铠甲、神兽坐骑和生活用具是本文首次识别的命名实体类型，目前没有可以对比的系统。以上是基于不同模型和数据规模等的对比结果，对识别方法或模型的有效性无法做出明确的评价。

为了验证本文所提出的基于藏文音节和深度

学习结合的格萨尔史诗命名实体识别方法的有效性，从格萨尔史诗（以四大降魔史为主）中随机抽取 98918 条句子作为训练集和 9289 条句子作为测试集进行实验。以及同一数据集上设置了基于藏文音节为基本单元的 5 组对比试验，对比结果如表 3 所示。

表 3　格萨尔史诗命名实体识别实验对比

Table3 Experimental Comparison of GesarNER

模型	P	R	F1
TS-GRU-CRF	85.90	81.62	83.56
TS-IDCNN-CRF	87.62	84.16	85.89
TS-LSTM	83.55	80.23	81.89
TS-LSTM-CRF	85.29	83.17	84.23
TS-BILSTM-CRF	**96.87**	**97.11**	**96.99**

实验表明，由于双向的 LSTM 能够获取上下文有效信息特征，再加上 CRF 能够充分考虑标注序列的顺序性，得到全局最优标注序列。相比于其他四种模型方法，基于 TS-BILSTM-CRF 模型的命名实体识别的 P（Precision）、R（Recall）和 F1 值（F-Score）三项指标分别提升了 9.25 个百分点、12.95 个百分点和 11.1 个百分点。具体识别效果如表 4 所示。

表 4　格萨尔史诗命名实体识别实验结果

Table4 Experimental Results of GesarNER

NE	TS-BILSTM-CRF		
	Precision	Recall	F-Score
ALL	**96.87**	**97.11**	**96.99**
PER	96.50	97.32	**96.91**
LOC	97.34	97.35	**97.34**
ORG	98.22	96.91	**97.56**
WEA	94.56	93.83	**94.19**
HER	95.02	96.09	**95.55**
LIV	98.37	96.37	**97.36**

（3）为了进一步展现实验的识别效果，本文设计并开发了格萨尔史诗命名实体识别系统，主要包括命名实体识别、实体种类分析和实体出现次数统计等功能。系统可视化界面如图 6 所示。

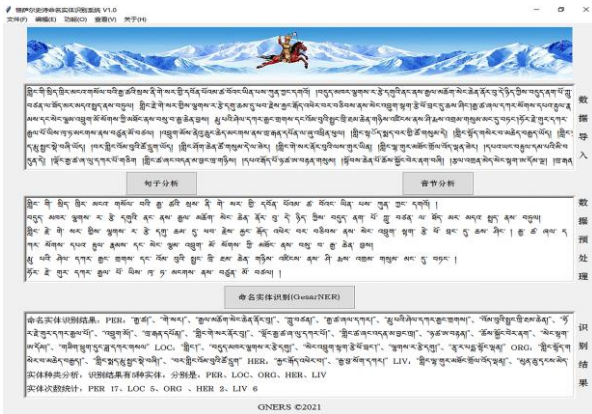

图 6　GesarNERS 可视化系统

Figure 6 GesarNERS Visualization System

6 结语

本文针对特定的格萨尔领域存在标注数量较少且实体识别困难、识别精度不高的问题，提出了一种基于 TS-BILSTM-CRF 的命名实体识别方法，并在自建数据集上实验验证，基于 TS-BILSTM-CRF 方法充分学习了文本的特征信息，使得三项指标均有较大程度的提升，优于本实验的其它模型。同时，本文的方法为其它专业领域的实体识别提供一种有效的解决思路。

下一步将会继续听取前辈们的意见和学习已有研究的基础上，更认真地分析和改进未能正确识别的原因及方法。加强对深度学习等新技术和新方法的学习，从六种类型的格萨尔史诗命名实体识别扩展到完整的藏文命名实体识别，为创建格萨尔知识图谱和藏文实体抽取等奠定坚实的基础。

参考文献

[1] 多拉,扎西加.词汇计量与史诗诸要素的解析——以语料库方法解构格萨尔史诗之《霍岭》[J].西藏大学学报(社会科学版),2014,29(03):103-110.

[2] ZHAO S,CAI Z P,CHEN H W,WANG Y,LIU F,LIU A F. Adversarial training based lattice LSTM for Chinese 2clinical named entity recognition[J].Journal of Biomedical Informatics,2019,99(14).103290.

[3] YAO L G,HUANG H S ,WANG K W,CHEN S H,XIONG Q Q. Fine-Grained Mechanical Chinese Named Entity Recognition Based on ALBERT-AttBiLSTM-CRF and Transfer Learning[J]. Symmetry,2020,12(12).1986.

[4] SUDHAKARAN G,MANJULA D,VIJAYAN S. Character level and word level embedding with bidirectional LSTM – Dynamic recurrent neural network for biomedical named entity recognition from literature[J]. Journal of Biomedical Informatics,2020,112(prepublish).103609.

[5] JUN K,ZHANG L X,JIANG M,LIU T S. Incorporating mul-ti-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics,2021,116.103737.

[6] DEBORA N,PIKAKSHI M,ELISABETTA F,MATTEO P,ENZA M. LearningToAdapt with word embeddings: Domain adaptation of Named Entity Recognition systems[J]. Information Processing and Management,2021,58(3).102537.

[7] 胡滨,耿天玉,邓赓,段磊.基于知识蒸馏的高效生物医学命名实体识别模型[J/OL].清华大学学报(自然科学版):1-7[2021-04-10].https://doi.org/10.16511/j.cnki.qhdxxb.2020.26.035.

[8] 李天然,刘明童,张玉洁,徐金安,陈钰枫.基于深度学习的实体链接方法研究[J/OL].北京大学学报(自然科学版):1-9[2021-04-10].https://doi.org/10.13209/j.0479-8023.2020.077.

[9] 加羊吉.基于语料库的藏语命名实体识别研究[D].西北

民族大学,2014.
[10] 华却才让,姜文斌,赵海兴,刘群.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.
[11] 王志娟,刘飞飞,赵小兵,宋伟.基于置信度的藏文人名识别的主动学习模型研究[J].中文信息学报,2019,33(08):53-59.
[12] 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.
[13] 降边嘉措.《格萨尔》大辞典[M].北京：海豚出版社出版，2017.3：289-290.
[14] STONEY C,ROBBINS R A,MCKONE E. A stimulus set of people famous to current generation Australian undergraduates, with recognition norms for face images and names[J]. Australian Journal of Psychology,2020,72(4).
[15] 降边嘉措.《格萨尔》大辞典[M].北京：海豚出版社出版,2017.3：48-49.
[16] 李向明.谈地名标准化与地名文化保护——以呆鹰岭为例[J].中国地名,2018(09):9-10.
[17] WUMAIER A,XU C Y,KADEER Z,LIU W Q,SAIMAI A. A Neural-Network-Based Approach to Chinese–Uyghur Organization Name Translation[J]. Information,2020,11(10):492.
[18] 才智杰.藏文词向量表示关键技术研究[D].青海师范大学,2018.
[19] 多拉,扎西加.藏文规范音节频率[M].中国社会科学出版社,2015,1-2.
[20] ALEX G,JURGENS. FramewisePhonemeClassificationwithBidirectional LSTM and Other Neural Network Architectures[J].Neural Networks,2005,18(05):602-610.
[21] 尹学振,赵慧,赵俊保,姚婉薇,黄泽林.多神经网络协作的军事领域命名实体识别[J].清华大学学报(自然科学版),2020,60(08):648-655.

Research on Gesar Epic Named Entity Recognition Based on Deep Learning

HUAN Keyou[1,2,3*] , HUAQUE Cairang[1,2,3] , CAIRANG Dangzhi[1,2,3] ,DUOJIE Cairang[1,2,3]

(1. College of Computer ,Qinghai Normal University, Xining 810006 , China;
2. The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810008 , China;3. Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province, Xining, Qinghai 810008 ,China)

Abstract: In order to make a profound research on the basic structure, classification method and automatic recognition technology of Tibetan named entity, this thesis tries to carry out a basic research in the field of information processing, which also aims to improve the lexical analysis, syntactic analysis and The article semantic analysis of Tibetan, as well as machine translation. Therefore, it is devoted to the research of Gesar is epic, which has various entity types, and put forwards six types of named entities. At the same time, the Gesar is epic named entity recognition method based on Tibetan syllables and deep learning has also been proposed. With the help of the named entity practice book and test book with over 100,000 sentences of manual annotation, it is found that the accuracy rate, recall rate and f-value of named entity recognition are 96.87%, 97.11% and 96.99% respectively through experiment. It has not only meet the application requirements of studying Gesar is epic named entity recognition, but also provides the basis for making Gesar is knowledge graph and the annotation standards of Tibetan named entity.

Key words: Gesar; named entity recognition; deep learning