DOI: 10.3979/j.issn.1673-825X.2020.04.018

基于 BiLSTM_CRF 模型的藏文分词方法

王莉莉[1],王宏渊[1],白玛曲珍[1],杨鸿武[1,2,3]
(1.西北师范大学 物理与电子工程学院,兰州 730070;2.甘肃省智能信息技术与应用工程研究中心,兰州 730070;
3.互联网教育数据学习分析技术国家地方联合工程实验室,兰州 730070)

摘　要: 藏文分词是实现藏文语音合成和藏文语音识别的关键技术之一。提出一种基于双向长短时记忆网络加条件随机场(bidirectional long-short-term memory with conditional random field model,BiLSTM_CRF)模型的藏文分词方法。对手工分词的语料经过词向量训练后输入到双向长短时记忆网络(bidirectional long-short-term memory,BiLSTM)中,将前向长短时记忆网络(long-short-term memory,LSTM)和后向 LSTM 学习到的过去输入特征和未来输入特征相加,传入到线性层和 softmax 层进行非线性操作得到粗预测信息,再利用条件随机场(conditional random field,CRF)模型进行约束性修正,得到一个利用词向量和 CRF 模型优化的藏文分词模型。实验结果表明,基于 BiLSTM_CRF 模型的藏文分词方法可取得较好的分词效果,分词准确率可达 94.33%,召回率为 93.89%,*F* 值为 94.11%。

关键词: 文本分词;长短时计忆网络;深度神经网络;词向量;民族语言
中图分类号: TP391.1;TN912.33　　**文献标志码:** A　　**文章编号:** 1673-825X(2020)04-0648-07

Tibetan word segmentation method based on BiLSTM_CRF model

WANG Lili[1], WANG Hongyuan[1], BAIMA Quzhen[1], YANG Hongwu[1,2,3]
(1. College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, P.R. China;
2. Engineering Research Center of Gansu Province for Intelligent Information Technology and Application, Lanzhou 730070, P.R. China;
3. National and Local Joint Engineering Laboratory of Data Learning and Analysis Technology for Internet Education, Lanzhou 730070, P.R. China)

Abstract: Tibetan word segmentation is one of the key technologies to realize Tibetan speech synthesis and Tibetan speech recognition. This paper proposes a Tibetan word segmentation method based on bidirectional long-short-term memory with conditional random field (BiLSTM_CRF) model. Firstly, the corpus of manual word segmentation is input into BiLSTM model after word vector training. Then the past input features acquired by forward long-short-term memory network (LSTM) are added with the future input features acquired by backward LSTM. The nonlinear operation is carried out in the linear layer and the softmax layer to obtain the rough prediction information. The constraint correction is finally carried out in the conditional random field (CRF) model to obtain a Tibetan word segmentation model optimized by word vector and CRF model. The experimental results show that the proposed method can achieves 94.33% on word segmentation accuracy, 93.89% on recall rate and 94.11% on *F* value.

Keywords: text segmentation; long-short-term memory network; deep neural network; word vector; ethnic language

收稿日期: 2018-12-13　**修订日期:** 2020-03-03　**通讯作者:** 杨鸿武　yanghw@nwnu.edu.cn
基金项目: 国家自然科学基金(11664036,61263036);甘肃省高等学校科技创新团队项目(2017C-03)
Foundation Items: The National Natural Science Foundation of China(11664036, 61263036); The High School Science and Technology Innovation Team Project of Gansu(2017C-03)

0 引　言

藏语是我国一种历史悠久的民族语言，其使用范围遍及西藏、青海、甘肃、四川、云南等西部地区以及尼泊尔、不丹、巴基斯坦、印度等国家的部分地区，使用人口多达 800 万，分布地域广大，传承和记载了丰富多彩的藏民族文化。藏语属汉藏语系藏缅语支，与汉语拼音一样，也是一种拼音文字，通过藏语拼音进行拼写，且字母组合排序有严格的规则，需按照从左到右，从上到下的顺序进行书写，如图 1。其中基字作为每个音节的核心位置，用来确定该音节的中心辅音位置。基字分别与前加字、上加字、下加字组合起来形成藏字的声母，元音与后加字和再后加字组合成藏字的韵母。藏文各音节之间由音节点分隔，但是词与词之间却没有分隔标记。同汉语类似，计算机的所有语言知识都来自机器词典、句法规则以及有关词和句子的语义、语境、语用知识库等。在藏文信息处理中，只要涉及到句法、语义，就需要以词为单位进行处理，如藏文信息检索、文语转换、文本校正、机器翻译、文本分类、自动摘要等。因而在藏文信息处理中首先需要解决词的切分问题。

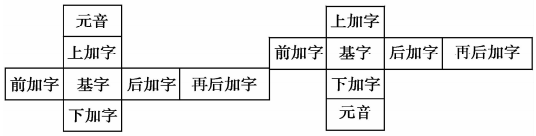

图 1　藏语音节的组成结构

Fig.1　Structure of the tibetan syllables

目前，藏文分词的主流方法有 3 种，分别是基于规则的方法、基于统计的方法以及规则与统计相结合的方法。

基于规则的方法主要是利用词典、格助词和虚词等规则[1]，采用一些规则算法，如最大匹配算法[2]，双向扫描匹配法[3]，逐次匹配法[4]等，选择出最佳匹配词作为分词结果[5]。虽然利用规则能实现藏文分词，但是由于藏文的复杂性以及紧缩词和未登录词的存在，使得基于规则的方法很难适应大规模文本语料分词。

基于统计的方法，主要是通过训练统计模型，输出概率最大的结果作为藏文分词的结果。文献 [6] 利用隐马尔科夫模型，将汉语分词系统 Segtag 移植到藏文分词系统上，实现了藏文分词。文献 [7] 采用基于条件随机场模型的 4 字位标注集进行藏文分词。文献 [8] 采用基于条件随机场模型的 6 字位标注集进行藏文分词。但是基于统计的方法会自动分出一些出现频率高、但并不是词的常用字组，并且对常用词的识别精度差，对大规模语料时空开销大。

基于规则与统计相结合的方法主要是在统计模型的基础上加一些规则对模型以及模型训练结果进行修正。文献 [9] 在基于条件随机场模型的基础上，利用藏文中紧缩词的识别与分词模块相结合的方法，进一步提高了藏文分词的效果。文献 [10] 使用基于词位切分的条件随机场模型，将藏语黏写形式的规则特征融合到藏文分词研究中，提高了藏文分词准确率。文献 [11] 采用基于知识融合的条件随机场方法，对条件随机场模型分词结果进行分析，总结规则，并用规则对分词结果进行修正，实现规则与统计相结合，但在构建规则的过程中往往需要大量的藏语语言学知识，需要处理规则之间的冲突问题，而且构建规则的过程费时费力、可移植性不好。

近年来，深度学习的方法也应用到了藏文分词中[12]。但是目前的研究，并没有充分挖掘神经网络任务中藏文句子上下文相关信息对藏文分词准确性的影响。

对于时序数据，双向长短时记忆（bidirectional long short-term memory，BiLSTM）网络[13-14]可以融合 2 组学习方向相反的长短时记忆（long short-term memory，LSTM）层，能使得当前词即包含历史信息，又包含未来信息，更有利于对当前词进行标注。同时为了使输出结果更优化，通过后接条件随机场（conditional random field，CRF）层，在整个序列上学习最优的标签序列，得到最优结果。本文提出一种结合 BiLSTM 和 CRF 的基于双向长短时记忆加条件随机场（bidirectional long short-term memory with conditional random field，BiLSTM_CRF）模型的藏文分词方法，将藏文文本表示为词向量，进行模型训练，能够提高藏文分词的准确率。

1 基于 BiLSTM_CRF 模型的藏文分词框架

基于 BiLSTM_CRF 的藏文分词方法主要由 4 部分构成，如图 2。第 1 部分是藏文语料预处理；第 2 部分是将手工分词的语料转换为词向量矩阵；第 3 部分是将词向量矩阵输入到 BiLSTM 模型，利用前向 LSTM 和后向 LSTM 的过去输入特征和未来输入特征相加，传入到线性层和 softmax 层进行非线性操作，得到粗预测信息；第 4 部分是将粗预测信息传入

给 CRF 模型进行约束性修正,得到一个利用词向量和 CRF 模型优化的藏文分词模型,同时利用 Dropout 网络[15]对 BiLSTM_CRF 模型的训练时间和过拟合问题进行优化处理。

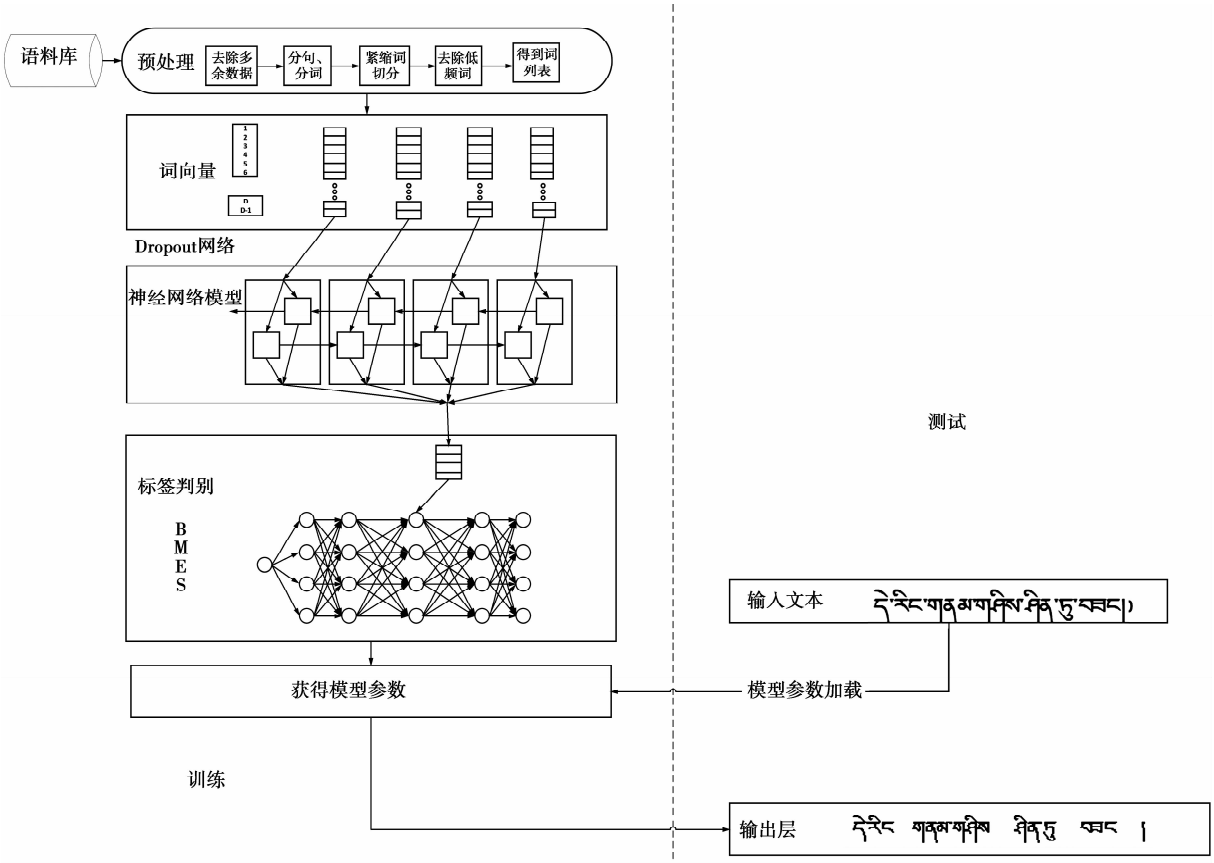

图 2 BiLSTM_CRF 模型框架

Fig.2 Framework of BiLSTM_CRF model

1.1 预处理

首先对训练语料进行预处理,根据藏文单垂符,云头符等标点符号,对藏文训练语料进行分句。然后利用词典,采用逆向最大匹配算法[16]对训练语料进行粗略分词,并对预分词结果进行人工校对,对紧缩词进行全切分,用确定的标志符替换藏文习语、连续中英文字符和字等。如用<NUM>表示数字,用<IDIOM>表示藏文习语,用<TOKEN>替换英文、中文等其他语言字符。最后将处理后的训练语料中的藏词按照出现频率高低进行排序,去除低频词,为后续词嵌入提供基础。

1.2 词向量

本文利用 Word2vec 工具包中的连续词袋(continuous bag of words, CBOW) 模型构建词向量空间模型,在已知当前词 $W_{(t)}$ 的前向词 $W_{(t-2)}$,$W_{(t-1)}$ 和后向词 $W_{(t+1)}$,$W_{(t+2)}$ 的前提下预测当前词 $W_{(t)}$。CBOW 模型主要由输入层、投影层和输出层构成,如图 3。

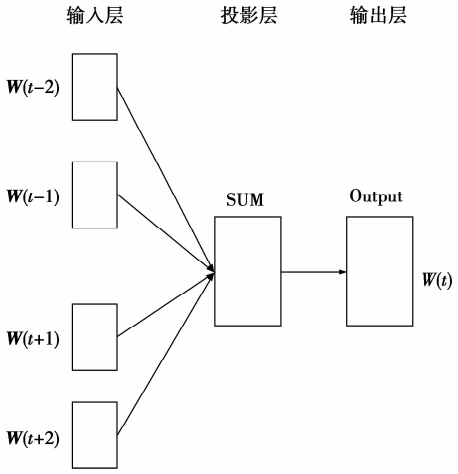

图 3 CBOW 模型

Fig.3 CBOW model

输入层: 将语料库中预处理后的藏词 $C_{(t-m)}$,$C_{(t-m+1)}$,…,$C_{(t+m)}$,通过 CBOW 模型窗口顺序读取,然后由哈希表得到投影层相应词位置 $W_{(t-m)}$,$W_{(t-m+1)}$,…,$W_{(t+m)}$,获得当前词 $W_{(t)}$ 的 m 个上下文

相关词 $context(W(t))$。

投影层: 对 m 个词的 $context(W(t))$ 做累加求和,表示为

$$V(t)=\sum_{t-n}^{t+n} context(W(t)) \tag{1}$$

输出层: 根据当前词 $W_{(t)}$ 的上下文相关信息生成该词的向量值,表示为

$$P(W(t)\mid context(W(t)))=\prod_{t-n}^{t+n} f(V(t),\theta) \tag{2}$$

$$f(V(t),\theta)=\frac{1}{1+e^{-V(t)\theta x}} \tag{3}$$

$f(V(t),\theta)$ 表示一个结点被分为正类的概率。

我们以 $W_{(t)}$ 作为预测单词,去掉中间隐藏层,简化了神经网络模型,且用低维实数向量表示藏词,实现藏词向量化。该特征向量可以刻画词与词在语义和语法上的相关性,即语义相近的单词向量之间的距离也会较近,而且还能够很好地表示词的语义和句法信息。此外,周围词的位置不会影响预测结果。同时,训练好的词向量输入到 BiLSTM_CRF 模型之前设置 Dropout 网络以缓解过拟合。

1.3　BiLSTM 模型

长短时记忆网络[17]是一种时间递归神经网络,可以很好地对长距离依赖信息进行建模,模型结构如图 4。

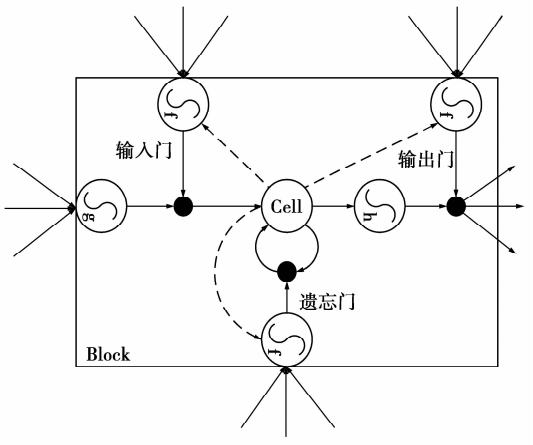

图 4　LSTM 网络结构

Fig.4　Network structure of LSTM

信息从输入门输入,通过循环连接的细胞单元,该单元用于控制流向输入门的信息和控制遗忘门之前的细胞状态的遗忘门。每个时刻各单元计算式表示为

$$i_t=\sigma(w_i\cdot[h_{t-1},x_t]+b_i]) \tag{4}$$

$$f_t=\sigma(w_f\cdot[h_{t-1},x_t]+b_f) \tag{5}$$

$$o_t=\sigma(w_o\cdot[h_{t-1},x_t]+b_o) \tag{6}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot(\tanh(w_c\cdot[h_{t-1},x_t]+b_c)) \tag{7}$$

$$h_t=o_t\odot\tanh(c_t) \tag{8}$$

(4)—(8) 式中: i_t,f_t,o_t,c_t 分别表示 t 时刻输入门,忘记门,输出门和细胞状态的输出; h_t 和 x_t 表示 t 时刻的隐藏层向量和输入向量; σ 表示 sigmoid 激活函数,输出 [0,1] 内的数值,描述每个部分有多少量可以通过,0 代表“不许任何量通过”,1 代表“允许任意量通过”; w 和 b 分别表示权重矩阵和偏置向量。

尽管 LSTM 网络在藏文分词中有不错的表现,但该模型是从左向右推进的,导致句子中前面词的权重比后面词的权重小,而对于藏文分词而言,句子中每个词的权重应该相同。因此,为了更好地获得藏词的前后上下文信息,本文使用 BiLSTM 网络进行建模。该模型结合了前向 LSTM 和后向 LSTM 模型,除了使用过去输入特征和语句级标记信息之外,也可以使用未来的输入特征。将 BiLSTM 模型学习到的特征输入到 softmax 中预测藏词位置信息。本文采用 4 词位标注集(B,M,E,S) 进行标注[18],B(Begin) 标注藏词的开始,M(Middle) 标注藏字在藏词的中间位置,E(End) 标注藏词的结束,S(Single) 则标注单字符词的藏文,所有藏文字符位置信息用分数表示,选择每个藏文字符得分最高的标签输出,从而判断位置。同时,对于 BiLSTM_模型,使用反向传播算法来训练该网络,更新参数信息 α 表示为

$$\alpha\leftarrow\alpha+\gamma\frac{\partial ln\rho(y\mid x,\alpha)}{\partial\alpha} \tag{9}$$

(9) 式中: x 为输入向量; ∂为输入向量对应的标记; ρ 为 softmax 计算的最大概率。通过随机梯度下降算法(stochastic gradient descent,SGD)[19]估计 α 值。

1.4　CRF 模型

CRF 层作为 BiLSTM 模型外层解码结构,主要作用是对 softmax 预测的藏文字符位置信息进行修正。虽然 BiLSTM 能够学习上下文的信息,但是输出结果之间相互独立,softmax 分类器只是在每一步挑选一个最大概率值的标签输出,会导致在一个藏词中出现 BBME 的位置标签。在实际情况中,B 后面不可能是 B,故 softmax 分类器后接 CRF 模型进行句子级的序列标注。

CRF 层的输入参数是 BiLSTM 层的输出结果,一个 $n\times m$ 的矩阵 p,其中 n 是藏词个数,m 是标签种

类,定义 p_{ij} 是从第 i 个标签到第 j 个标签的转移分矩阵,对于一个长度等于句子长度的预测标签序列 $\boldsymbol{y}=(y_1,y_2,\cdots y_n)$,它的概率为

$$score(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{n}\boldsymbol{p}_{i,\boldsymbol{y}_i}+\sum_{i=1}^{n+1}\boldsymbol{A}_{\boldsymbol{y}_{i-1},\boldsymbol{y}_1} \tag{10}$$

(10) 式中,y_1、y_n 是预测句子的结束和起始标记。整个序列概率由 2 部分构成,一部分是 BiLSTM 层输出的 $\boldsymbol{p}$ 矩阵,另一部分是 CRF 层的转移矩阵 $\boldsymbol{A}$。

CRF 层训练时对一个训练样本 $\boldsymbol{x},\boldsymbol{y}^x$ 最大化标记的似然函数为

$$\ln(p(\boldsymbol{y}^x\mid\boldsymbol{x}))=score(\boldsymbol{x},\boldsymbol{y}^x)-\ln(\sum_{y'}e^{score(\boldsymbol{x},\boldsymbol{y}')}) \tag{11}$$

(11) 式中,y' 表示真实的标记值。在预测过程时使用动态规划的 Viterbi 算法来求解最优路径为

$$\boldsymbol{y}^*=\arg\max_{y'}score(\boldsymbol{x},\boldsymbol{y}') \tag{12}$$

2 藏语分词实验

2.1 实验数据

目前,还没有公开的藏语分词语料库。为验证本文提出的 BiLSTM_CRF 模型在藏文分词中的有效性,实验语料从中国西藏藏文网,青海藏文网,康巴卫视等网站手工进行摘录,主要包括各类新闻,名人轶事,小说等题材,共 3 000 篇藏文文章。之后经过分句、最大匹配法分词,人工对文本进行过滤、校对,最终得到 125 386 句藏文语料,共计 1 196 907 个藏词。实验中随机选取 80% 的句子作为训练集,10% 的句子作为验证集,余下 10%的句子作为测试集。

2.2 BiLSTM_CRF 模型参数设置

BiLSTM_CRF 模型参数配置如下。

词向量训练部分,窗口大小设为 5,即以当前藏词为基准,前后各取 5 个藏词,如果不够 10 个藏词,添加 0 向量作为补充,如果超过 10 个藏词,添加 1 向量作为补充。在训练时,通过改变词向量维度来提高藏文分词的速度和准确度。

对 BiLSTM_CRF 模型采用反向传播算法进行训练,每次更新一个训练样本参数,并使用随机梯度下降算法,学习率设为 0.01,梯度裁剪为 5.0。

BiLSTM_CRF 模型前向和后向各拥有一个 LSTM 层,其尺寸设置为 100。实验表明,调节此维度对藏文分词准确率没有显著影响。下面的实验通过调优词向量维度寻找最优参数,而且改变不同的 Dropout 比例,看其对实验结果和训练时长的影响,以及测试不同实验模型对藏文分词准确率的影响。

2.3 实验设计及结果分析

2.3.1 词向量维度设计

词向量的维度对藏文分词的速度和准确度都起着至关重要的作用。在保证 BiLSTM_CRF 模型其他参数不变的情况下,使用准确率(precision,P),召回率(recall,R)和综合 F 值对模型进行评估。不同维度词向量对分词结果的影响如图 5。

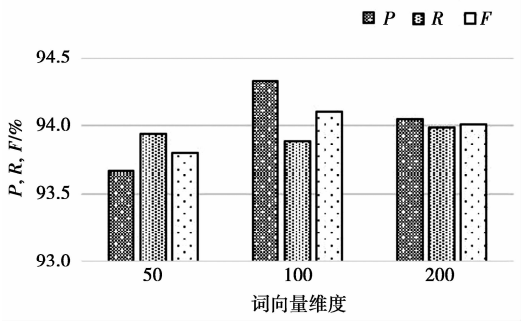

图 5 不同词向量维度下藏文分词的 P、R、F 值对比

Fig.5 Comparison of P, R and F values of tibetan segmentation in different vector dimensions

由图 5 可知,当词向量维度为 100 时 BiLSTM_CRF 模型性能相对最优,其 R 值和 F 值高于维度为 50 和 200 的 R 值和 F 值。因为词向量维度取 50 时,不能完全覆盖句子的特征,效果不优。词向量维度大于 200 时,模型训练时间加长且性能下降,说明在藏文分词时,词向量维度不宜过小或过大。

2.3.2 Dropout 比例设计

Dropout 网络可以防止深度神经网络过拟合。在词向量维度为 100 时,采用同样大小的语料对不同比例的 Dropout 进行实验。由表 1 实验结果可知,当 Dropout 比例为 0%时,模型容易过拟合;Dropout 比例为 20% 时,性能相对最优,F 值达到 94.11%;当 Dropout 比例为 70%时,F 值显著降低。Dropout 比例越高,训练网络中不工作的节点越多,虽然训练时间减少但是模型欠拟合,导致性能下降。在语料库规模相对不大时,更高比例的 Dropout 网络会对实验结果产生负面影响。

2.3.3 BiLSTM_CRF 模型实验结果的对比分析

为了对比 Dropout 比例为 20%,词向量维度为 100 时的 BiLSTM_CRF 模型与 CRF,LSTM,BiLSTM,LSTM_CRF 模型在藏文分词上的性能,本文采用相同的实验参数、相同的实验数据进行藏文分词,实验结果如表 2。

表 1　Dropout 比例

Tab.1　Dropout ratio

Dropout 比例	*P*/%	*R*/%	*F*/%
0%	92.06	91.26	91.66
20%	94.33	93.89	94.11
50%	93.39	92.97	93.18
70%	89.43	90.12	89.77

表 2　不同方法分词结果对比

Tab.2　Comparison of word segmentation results by different methods

实验设计	训练时间/h	*P*/%	*R*/%	*F*/%
CRF	12	89.97	91.01	90.49
LSTM	7	90.12	91.74	90.92
BiLSTM	8	93.12	92.49	92.80
LSTM_CRF	9.5	93.43	92.56	92.99
BiLSTM_CRF	11.5	94.33	93.89	94.11

1) 通过对比表 2 中第 1 行数据和其他 4 行数据，可知基于深度学习的 LSTM，BiLSTM，LSTM_CRF 和 BiLSTM_CRF 模型的实验结果均比传统的 CRF 模型的藏文分词效果好，*F* 值依次提高 0.43%，2.31%，2.5%，3.62%，且运行时间少。因此，在大规模无相关藏语语言规则的情况下基于深度学习的模型在藏文分词上效果更好。

2) 由表 2 可知，在藏文分词中 BiLSTM 模型比 LSTM 模型的分词准确率更高。虽然训练时间长，但是对同一句藏文进行分词测试，二者分词速度差不多。在同样的语料库下，BiLSTM 模型的精确度为 93.12%，召回率为 92.49%，*F* 值为 92.80%，其精确率、召回率、*F* 值比 LSTM 模型分别提高了 3%，0.75%，1.88%。同时，BiLSTM_CRF 模型与 LSTM_CRF 模型相比，其精确率、召回率、*F* 值分别提高了 0.9%，1.33%，1.12%，充分说明 BiLSTM 模型比 LSTM 模型更能提高藏语分词的准确率。

3) BiLSTM_CRF 模型中 CRF 层的作用由表 2 可知，LSTM_CRF 模型相比 LSTM 模型，*F* 值提高了 2.07%，BiLSTM_CRF 模型相比 BiLSTM 模型，其精确率、召回率、*F* 值分别提高了 1.21%，1.4%，1.31%，可以知道 CRF 模型作为 LSTM，BiLSTM 模型的最后一层，可以对 softmax 预测的结果进行修正，输出全局最优化的藏文分词结果。虽然增加 CRF 层，训练时间变长，但并不影响测试时间。

为进一步说明 BiLSTM_CRF 模型在藏文分词中的适应性，本文将该模型分词效果与其他学者所提出的藏语分词方法进行了比较，说明本文所提方法在藏文分词中的适用性较好。虽然现有藏文分词标注系统的方法、语料库和词典各不相同，没有统一规范的语料库，很难进行严格的比较，但在本文设计的语料库上，将基于 BiLSTM_CRF 模型的藏文分词与正向最大匹配、逆向最大匹配、双向最大匹配、条件随机场以及最大熵模型的方法进行比较，其准确率有明显提高。而且基于 BiLSTM_CRF 模型的藏文分词方法并未使用其他语言规则进行辅助。

3　结束语

本文将 BiLSTM_CRF 模型应用到藏文分词中，在输入层输入 CBOW 模型生成的藏文词向量，经过中间层特征学习和 softmax 分类得到粗预测结果，最后在输出层采用 CRF 模型对粗预测结果进行约束性修正，得到 Dropout 比例为 20%，词向量维度为 100 的最佳分词效果的 BiLSTM_CRF 模型。实验结果表明，该方法能够有效提高藏文分词的准确率。

参考文献:

[1] 祁坤钰.信息处理用藏文自动分词研究 [J].西北民族大学学报(哲学社会科学版)，2006(4):92-97.
QI K Y. Research on Tibetan Automatic Word Segmentation for Information Processing [J]. Journal of northwest university for nationalities, 2006(4): 92-97.

[2] 刘汇丹，诺明花，赵维纳，等.SegT: 一个实用的藏文分词系统 [J].中文信息学报，2012，26(1):97-104.
LIU H D, NUO M H, ZHA0 W N, et al. SegT: A Practical Tibetan Word Segmentation System [J]. Journal of Chinese information processing, 2012, 26(1): 97-104.

[3] SUN Y, YAN X D, ZHAO X B, et al. Research on Some Key Technologies of Tibetan Automatic Word Segmentation [C] //2011 4th International Conference on Intelligent Networks and Intelligent Systems. Kunming, China: IEEE, 2011: 188-191.

[4] 才华.基于小字符集的藏文自动分词技术研究 [J].西藏大学学报(自然科学版)，2013，28(2): 43-47.
CAI H. Research on Tibetan Automatic Word Segmentation Technology Based on Small Character Set [J]. Journal of Tibet university (Natural Science Edition), 2013, 28(2): 43-47.

[5] WEI X, PENG J. design of automatic Tibetan word segmentation system based on word frequency learning and dynamic word frequency updating [J]. Computer Applications & Software, 2014, 5(31): 106-109.

[6] 史晓东，卢亚军.央金藏文分词系统 [J].中文信息学

报, 2011, 25(4): 54-57.
SHI X D, LU Y J. A Tibetan Segmentation System—Yangjin [J]. Journal of Chinese information processing, 2011, 25(4): 54-57.

[7] JIANG T, YU H Z, JAM Y. Tibetan word segmentation system based on conditional random fields [C]//2011 IEEE 2nd International Conference on Software Engineering and Service Science. Beijing, China: IEEE, 2011: 446-448.

[8] 康才畯,龙从军,江荻.基于词位的藏文黏写形式的切分 [J].计算机工程与应用, 2014, 50(11): 218-222.
KANG C J, LONG C J, JIANG D. Segmentation of Tibetan abbreviated forms based on word position [J]. computer engineering and applications, 2014, 50(11): 218-222.

[9] 李亚超,江静,加羊吉,等.TIP-LAS:一个开源的藏文分词词性标注系统 [J].中文信息学报, 2015, 29(6): 203-207.
LI Y C, JIANG J, JIA Y J, et al. TIP-LAS: An Open Source Toolkit for Tibetan Word Segmentation and POS Tagging [J]. Journal of Chinese information processing, 2015, 29(6): 203-207.

[10] 康才畯.藏语分词与词性标注研究 [D].上海:上海师范大学, 2014.
KANG C J. Research on Tibetan Word Segmentation and Part of Speech Tagging [D]. Shanghai: Shanghai Normal University, 2014.

[11] 洛桑嘎登,杨媛媛,赵小兵.基于知识融合的 CRFs 藏文分词系统 [J].中文信息学报, 2015, 29(6): 213-219.
LUOSANG G D, YANG Y Y, ZHAO X B. Tibetan Automatic Word Segmentation Based on Conditional Random Fields and Knowledge Fusion [J]. Journal of Chinese information processing, 2015, 29(6): 213-219.

[12] 李博涵,刘汇丹,龙从军,等.基于深度学习的藏文分词方法 [J].计算机工程与设计, 2018, 39(1): 194-198.
LI B H, LIU H D, LONG C J, et al. Tibetan word segmentation based on deep learning [J]. Computer engineering and design, 2018, 39(1): 194-198.

[13] REN Y F, TENG C, LI F, et al. Relation classification via sequence features and bi-directional LSTMs [J]. Wuhan University Journal of Natural Sciences, 2017, 22(6): 489-497.

[14] 刘广峰,黄贤英,刘小洋,等. 基于主题注意力层次记忆网络的文档情感建模 [J]. 四川大学学报: 自然科学版, 2019, 56(5): 833-842.
LIU G F, HUANG X Y, LIU X Y, et al. Document sentiment modeling based on topic attention hierarchy memory network [J]. Journal of Sichuan University (Natural Science Edition), 2019, 56(5): 833-842.

[15] ELADEL A, EJBALI R, ZAIED M, et al, Fast deep neural network based on intelligent dropout and layer skipping [C]//2017 International Joint Conference on Neural Networks (IJCNN). Anchorage, AK, USA: IEEE, 2017: 897-902.

[16] 龙从军,刘汇丹.藏文自动分词的理论与方法研究 [M].北京:知识产权出版社,2016: 65-70.
LONG C J, LIU H D. Research on the Theory and Method of Tibetan Automatic Word Segmentation [M]. Beijing, China: Intellectual Property Publishing House, 2016: 65-70.

[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[18] KANG C, JIANG D, LONG C. Tibetan Word Segmentation Based on Word-Position Tagging [C]//2013 International Conference on Asian Language Processing. Urumqi, China: IEEE, 2013: 239-242.

[19] DIEDERIK P K, JIMMY L B. Adam: A method for stochastic optimization [EB/OL]. (2015-12-30) [2018-12-10].https://arxiv.org/pdf/1412.6980v8.pdf

作者简介:

王莉莉(1994—),女,甘肃兰州人,硕士研究生,主要研究方向为自然语言处理。E-mail: 1335502737@qq.com。

王宏渊(1980—),女,甘肃兰州人,硕士研究生,主要研究方向为自然语言处理。E-mail: 452303915@qq.com。

白玛曲珍(1997—),女,西藏山南市人,本科生,主要研究方向为自然语言处理。E-mail: 2493958310@qq.com。

杨鸿武(1969—),男,甘肃合作人,教授,博士,博士生导师,主要研究方向为自然语言处理、语音信号处理。E-mail: yanghw@nwnu.edu.cn。

(编辑: 陈文星)