第46卷 第10期 计 算 机 工 程 2020年10月
Vol. 46 No. 10 Computer Engineering October 2020

·开发研究与工程应用· 文章编号: 1000-3428(2020)10-0294-07 文献标志码: A 中图分类号: TP391.1

基于 Bi-LSTM 与 CRF 的泰语句子切分模型

李自荐,迟呈英,战学刚

(辽宁科技大学 计算机与软件工程学院,辽宁 鞍山 114031)

摘 要: 在自然语言处理领域中,对于泰语等东南亚语言的分句处理是一项具有挑战性的工作。将序列标注模型应用于句子切分任务,提出基于双向长短期记忆循环神经网络的句子边界自动识别模型。利用 Glove 词向量技术,将泰语句子中的词或字转换为不同维度的向量,进而将词或字向量组合成为句子向量输入模型进行训练。在此基础上,通过双向网络结构捕捉上下文信息以达到更好的句子切分效果。实验结果表明,该模型在泰语句子切分任务上表现出非常精准的识别效果。

关键词: 自然语言处理; 句子切分; 深度学习; 循环神经网络; 长短期记忆网络; 泰语

开放科学(资源服务)标志码(OSID):

中文引用格式: 李自荐,迟呈英,战学刚. 基于 Bi-LSTM 与 CRF 的泰语句子切分模型[J]. 计算机工程,2020,46(10):294-300.

英文引用格式: LI Zijian,CHI Chengying,ZHAN Xuegang. Thai sentence segmentation model based on Bi-LSTM and CRF[J]. Computer Engineering,2020,46(10):294-300.

Thai Sentence Segmentation Model Based on Bi-LSTM and CRF

LI Zijian,CHI Chengying,ZHAN Xuegang

(School of Computer Science and Software Engineering,
University of Science and Technology Liaoning,Anshan,Liaoning 114031,China)

【Abstract】 In the field of Natural Language Processing,clause processing of Southeast Asian languages such as Thai is a challenging task. Therefore,sequence tagging model is applied to sentence segmentation and a sentence boundary automatic recognition model based on bidirectional Long Short-Term Memory cycle neural network is proposed. The words or characters in Thai sentences are transformed into vectors with different dimensions by using Glove word vector technology,and then the word vectors or character vectors are combined into a sentence vector and are input into the model for training. On this basis,the context information is captured through the bidirectional network structure to achieve better sentence segmentation effect. The experimental results show that the model is very accurate in the task of sentence segmentation in Thai.

【Key words】 Natural Language Processing(NLP); sentence segmentation; deep learning; Recurrent Neural Networks(RNN); Long Short-Term Memory(LSTM) network; Thai

DOI: 10.19678/j.issn.1000-3428.0055669

0 概述

近年来,深度学习技术在机器翻译[1-3]、数据抽取[4]、情感分析[5-6]等自然语言处理(Natural Language Processing,NLP)领域得到了广泛应用并且取得了快速发展。在 NLP 领域的诸多子任务中,分句任务是一种较为常见的任务。目前此类研究大多基于汉语、英语等被广泛使用的语言[7-8],其主要原因是这些语言存有大量标准化的数据集,能够为深度学习模型提供充足且优质的训练语料。而在泰语等东南亚语种的研究工作中,可用的优质语料数量不足,许多工作的训练数据来自互联网上的大规模非结构化数据[9]。数据来源的多样性和数据质量的不一致性造成在获取大规模数据集的过程中,很有可能出现掺杂大量段落的超长句子,这将对神经网络模型的收敛速度和模型最终效果产生影响。此外,在许多分句任务中,用户输入的数据不是一句话,而是一篇文章或一个段落,此时需要句子切分工具将段落信

基金项目: 国家自然科学基金(61672138)。

作者简介: 李自荐(1995—),男,硕士,主研方向为自然语言处理、深度学习; 迟呈英(通信作者),教授; 战学刚,副教授、博士。

收稿日期: 2019-08-05 修回日期: 2019-10-11 E-mail: chichengying@ustl.edu.cn

息切分为单句信息,方便任务处理。由于语料库规模庞大,难以通过人工方式对语料库中的句子进行切分,因此需要利用计算机技术对其进行精确的自动化切分。

与马来语、缅甸语等东南亚语种类似,泰语在书写过程中,词与词、句与句之间并没有明显的分隔标记,如空格、标点符号等[10]。对于具有这种语言特性的语种而言,很难通过一些简单的规则完成句子的精确切分。就人类而言,对于自己比较擅长的语种,即使段落中的词句之间没有明显的句子边界标记,也能很容易地判断出句子边界的位置,这是因为人类不仅能够通过标点等标志进行判断,还能依靠句意信息的协助进行句子边界的识别。因此,让模型在学习过程中获得更多更有用的词意、句意信息是十分重要的。

由于深度学习方法在自然语言处理领域中强大的学习能力,本文将 Glove + Bi-LSTM + CRF 网络结构应用于泰语句子切分任务,提出基于双向长短期记忆(Long Short-Term Memory,LSTM)循环神经网络的句子边界自动识别模型,以快速准确地识别泰语段落中的句子边界信息。

1　相关知识

1.1　循环神经网络和双向循环神经网络

循环神经网络(Recurrent Neural Networks,RNN)是一种特殊的前向神经网络,该种网络结构能够使用一种定向循环策略对一组输入序列重复进行同样的操作。当需要处理的数据是存在依赖关系的序列时,循环神经网络能够非常好地处理这些存在依赖关系的序列,得到令人满意的结果。

长短期记忆网络是一种特殊的循环神经网络单元,该种网络单元不仅能够把当前时刻节点的状态作为下一时刻节点的输出,还可以选择性地保留此前时刻的部分信息,能够更好地利用序列内部的依赖关系。

在实际应用中,对于当前状态的判断,除了前面时刻的依赖信息,如果对后面时刻的信息进行部分保留并与前面时刻的信息共同作为条件,就能够通过多种状态信息的共同判断,提高当前时刻状态预测的准确率。为此,研究者们提出了双向循环神经网络(Bi-RNNs)。

1.2　分句技术

分句作为 NLP 领域中一项基础任务,在其所属领域一直承担着十分重要的角色并且受到广泛关注。在多数未经数据清洗的语料库中,句子的来源多种多样,既存在通过人工翻译方式得到的高质量双语语句对,也存在从互联网上利用网络爬虫技术自动爬取得到的语句对。若语料库中出现过多超长段落,将会影响模型的学习效果以及模型性能。因此,在每次处理大规模数据集或使用者的输入序列时,希望语料库中的句子均为单句,以便充分地利用全部数据。将呈段落形式出现的序列精确切分为单句是一项复杂的工作。目前,很多语种的分句技术已经达到极高的准确率,但是依然存在以下 3 个问题:

1) 许多语种使用的句子切分方法是基于规则的方法,需要经常维护且对新领域、新语种很难适应。

2) 许多句子切分方法虽然考虑到切分位置的上下文信息,但是并不能完整地保留句子的句意信息。

3) 对于一些具有特殊语言特性的语种并不能做到高精度的切分。

使用双向循环神经网络完成句子切分任务为解决以上问题提供了有效方法。

2　相关工作

以往的研究工作基于规则对句子进行切分,弊端是太过依赖人工编写规则[11-13]。机器学习技术兴起后,经常被用来进行自然语言处理领域的基础研究,创造了许多非常成功的方法,但是该技术在任务开始之前需要人工对文本进行大量的数据预处理以及特征抽取操作,使得模型的最终处理效果非常依赖前期的人工操作。随着深度学习技术的发展,该技术在自然语言处理的基础任务中取得了巨大突破。使用深度学习技术不仅使任务取得了很好的效果,还减少了模型对于特征的依赖,节约了人工成本。因此,深度学习技术成为目前自然语言处理领域中最重要的方法。

在以往的研究中,通常认为空格是分隔两句泰语的唯一符号,因此,泰语句子切分问题变成了一个分类消歧的问题[14-16]。事实上,根据泰国语言相关部门的规定以及在真实语料中观察的结果可知,空格字符可能出现在句子中的任何位置,而两句泰语之间并不一定会使用空格符号。泰国语言皇家协会规定,在某些词的上下文处需要使用空格进行分隔[17],包括:

1) 在感叹词和拟声词的上下文处使用空格进行分隔,例如โอ้(噢!)、อุ๊ย(哎呦!)。

2) 在连词前使用空格进行分隔,例如และ(和)、หรือ(或)。

3) 在表示数值的词前后使用空格,例如มีนักเรียน 20 คน(有 20 个学生),เวลา 10:00 น.(时间上午 10:00)。

然而,这些规定在实际使用中并没有得到严格遵守,不同的泰语使用者将空格添加到句中的位置不尽相同,这造成了在泰语语料中,空格符号可能出现在句子中的任何位置,而不仅仅出现在句末。据统计得知,在一个泰语语料库中,大约 23% 的句子末尾没有空格出现。图 1 展示了一个泰语句子结构,其中句末没有空格出现。

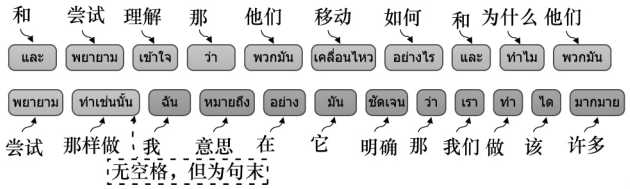

图 1 泰语句子结构

Fig. 1 Sentence structure of Thai

鉴于在泰语句子中空格问题的复杂性,文献[8]提出一种基于词的序列标记方法,将泰语句子切分问题看作词级的序列标记问题。该方法把句子中的空格看作普通单词进行处理,对句子中所有位置进行标记,根据标记结果判断句子切分位置。

3 本文模型

受文献[8]启发,本文将泰语句子切分任务看作是一个序列标注的问题,并且在切分过程中序列的长度是未知的。LSTM 神经单元作为一种特殊的神经网络单元,不仅能非常有效地处理变长的数据输入,还能解决在训练过程中出现的长距离依赖问题。基于双向长短期记忆(Bi-LSTM)和条件随机场(Conditional Random Field,CRF)[19-21]的泰语句子切分模型的词向量输入层、Bi-LSTM 网络层和神经网络输出层组成,模型框架如图 2 所示。

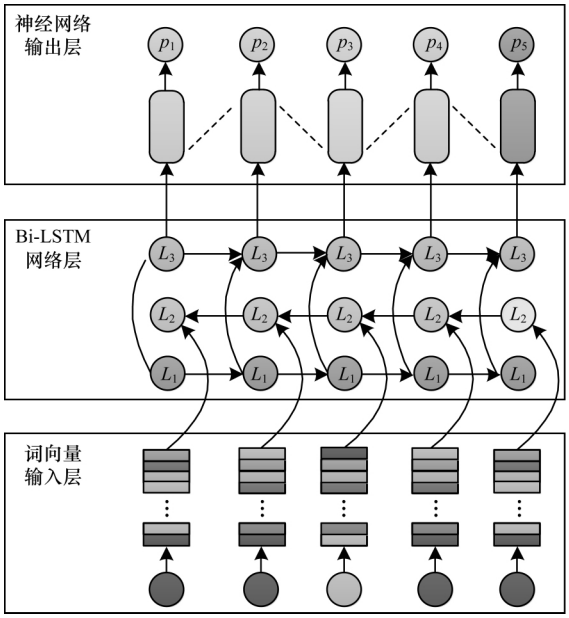

图 2 基于 Bi-LSTM 网络的泰语句子切分模型

Fig. 2 Thai sentence segmentation model based on Bi-LSTM network

3.1 词向量输入层

此前的研究已经证明,在深度学习研究中,使用词嵌入技术能够明显提升序列标注任务的准确率。本文使用 Glove 词嵌入技术,将句子中每个位置的词编码成指定维度的向量,将多个词向量组合得到的句子向量作为模型的输入,提供给模型学习。

Glove 是一种无监督的深度学习方法,将语料中的每一个词编码成不同维度的向量,转换后的向量能够保存被编码词的词意信息。Glove 通过计算两个词向量的欧式距离来调整合适的参数对词进行编码,通过该种方式使具有相似语意的词在向量空间上获得更近的距离。

使用 Glove 模型得到的单词编码不仅能表现出单个词的词意信息,还能表现出不同类别词之间的差异,这是因为使用欧式距离计算出的向量相似度是一个标量,可以对两个词之间的相似程度进行度量。但是单词往往能够表现出比标量更加复杂的相关程度,例如,“国王”和“王后”两个词都属于人类范畴,但无论从性别还是职位上来讲都存在一定差异。为了在最终结果中体现这些差异,Glove 捕获并计算很多向量之间的差异性,使得最终结果能够尽可能地表示两个不同词语之间的差异性。

3.2 Bi-LSTM 网络层

在切分模型的第 2 层,使用 Bi-LSTM 保证模型网络结构在训练过程中能够捕获不同单位之间的相互依赖关系,这种依赖关系主要包括前向和后向两种。通常情况下,一个循环神经网络在学习过程中除了接收一个序列在时刻 t 的输入向量 x_t 之外,还会接收来自前一时刻的状态信息 h_{t-1},h_{t-1} 与 x_t 共同决定 t 时刻的状态 h_t。将初始时刻状态 h_0 的值设置为 0。通过上述方法,循环神经网络可以在学习的过程中接收到来自其他时刻的信息并且有选择性地进行保留,这种方式能够使学习序列中连续元素共享相关信息,让模型得到更好的学习效果。

在本文模型中,为前向神经网络循环层输入一个编码向量 $[e_t]_{t=1}^{T}$,每一时刻的输出状态组成一个状态序列 $[h_t]_{t=1}^{T}$,该状态序列的长度为 $|e|$。在当前层,除了使用前向神经网络循环层之外,还通过在与正向 LSTM 相同的网络结构中输入顺序颠倒的字符序列实现了后向神经网络层,使得模型在学习过程中能够捕获当前时刻 t 的后向依赖关系。通过双向循环神经网络结构,在每一个时刻 t,可以得到两个方向的状态向量 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_{T+t-1}}$。

双向循环神经网络结构在训练过程中,通常使用随机梯度下降的方法优化网络结构中的参数,在每一步的训练过程中计算网络的整体误差,通过反向传播的形式迭代找到最优的权重,使得误差最小。然而,使用这种方式处理序列问题也存在一定弊端,即在网络训练过程中因为需要进行多次连续的相乘操作而使得参数调整量过小或过大,造成梯度消失或梯度爆炸,进而引起缩短误差的周期过长或权重的调整幅度过大,导致网络训练结果无法达

到最优。解决这一问题的主要方法是将循环神经网络结构中的普通单元替换成 LSTM 单元,该单元能够在每次的训练过程中忘掉一部分之前得到的状态信息,从而减少梯度消失或梯度爆炸问题的发生。

除此以外,模型还使用了 dropout 机制来增加泛化性能。在神经网络的学习过程中,如果网络结构的模型参数特别多,同时又缺少足够的训练样本,那么训练完成的模型很容易产生过拟合现象,即在模型训练样本上损失函数很小且准确率较高但在测试数据上损失函数很大且准确率较低。为此,使用 dropout 机制,在网络训练过程中,模型会随机使得某些神经元以一定的概率 p 停止工作,这样可以保证在每一批次的训练过程中,训练样本在不同的网络结构中进行学习,这使得模型的学习效果更好,不会依赖某些局部的特征。

3.3　神经网络输出层

在神经网络的输出层使用 CRF 层替代通用的 softmax 层,这是因为 LSTM 单元虽然能够处理长距离上下文信息,但是不能很好地处理标签之间的依赖关系,而 CRF 模型恰好能够弥补这一点。CRF 结构能够考虑相邻标签之间的依赖关系,得到的输出序列不是单位之间相互独立的,而是一个最优的输出序列 $s(x,y)$,表示如下:

$$s(x,y) = \sum_{i=0}^{n} \boldsymbol{A}_{y_i,y_{i+1}} + \sum_{i=0}^{n} \boldsymbol{p}_{i,y_i} \tag{1}$$

输出矩阵的大小为 $T \times k$,其中,T 为序列的长度,k 为标签的数量。$\boldsymbol{A}$ 为状态转移概率矩阵,$\boldsymbol{A}_{ij}$ 为从状态 i 转移到状态 j 所需要的状态转移概率,$\boldsymbol{p}_{i,y_i}$ 为当前位置表示内容的第 y_i 个标签的分数,定义如下:

$$\boldsymbol{p}_i = \boldsymbol{A}_s \boldsymbol{h}^{(t)} + \boldsymbol{b}_s \tag{2}$$

其中,$\boldsymbol{h}^{(t)}$ 是深度神经网络上一层 t 时刻输入数据 $\boldsymbol{x}_t$ 的隐藏状态,$\boldsymbol{A}_s$ 和 $\boldsymbol{b}_s$ 分别为状态转移矩阵和参数矩阵。

在模型中,输入部分是一个基于词编码的序列 $(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$,输出为概率向量 $(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n)$,其中 $\boldsymbol{y}_i$ 为 i 时刻对应标签的概率向量。

4　实验与结果分析

本文实验基于 IWSLT2015 公开数据集,对这一数据集中的所有单个句子进行切分并且验证模型的准确率。实验目标是探究 Bi-LSTM + CRF 模型在泰语句子切分任务中的应用效果,并且选取最合适的超参数以训练出最优的泰语分句模型。

4.1　数据集的选择与处理

本文的实验数据来自 IWSLT2015 公开数据集。经过筛选后,数据集中共包含单句泰语数据 191 538 条,均为拥有较高质量的泰语单语句子。各个数据集的数据量如表 1 所示。

表 1　数据集的数据量

Table 1　Data quantity of data set

数据集	数据量
训练集	187 538
校验集	2 000
测试集	2 000

对准备好的泰语数据进行以下处理,得到符合数据格式的泰语训练数据:

1) 样本构造。在训练集中所有泰语句子均为独立出现的单个句子,没有呈段落式出现的训练数据。由于模型需要呈段落式出现的泰语句子作为训练数据,为此采取如下策略: 将整个数据集中的句子全部打乱顺序并且逐条进行编码,每条句子都将得到一个唯一的编号,通过编号对指定的句子进行随机两两组合并且用 < sp > 标签标记生成的新句子的句末信息。该样本构造方法能够非常有效地构造出符合要求的泰语训练数据,同时增加样本的数量,如图 3 所示。

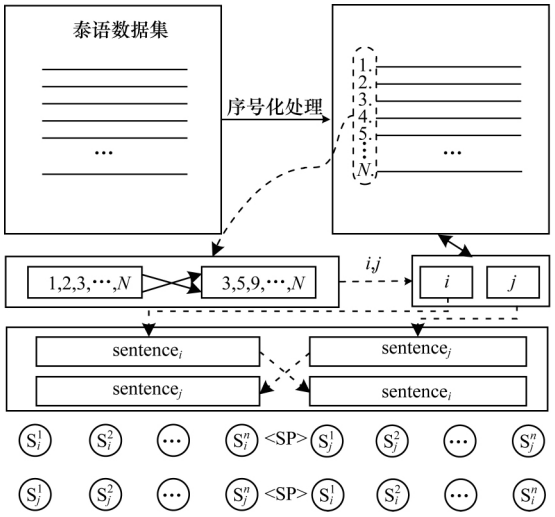

图 3　样本构造方法

Fig. 3　Sample construction method

2) 分词处理。泰语中词语之间并没有如英文一样的句子分隔标记,因此,需要对泰语数据进行分词处理。本文使用 SERTIS 泰语分词方法,该方法使用双向循环神经网络,按照字符级的切分方式,将每个字符转换成唯一的 id 提供给神经网络模型进行学习,同时使用 padding 以及 dropout 等方法,提升模型的切分效果和泛化性。最终模型在标准测试集上能够得到 99.11% 的 F1 值。

3) 特征抽取。在实验过程中,针对某些特定的任务需要对泰语训练数据进行相应的特征抽取,将抽取出来的特征提供给相关的模型进行训练。本实验对每个位置的上下文内容信息进行抽取,由当

前位置的内容及其上下文特征组合成一条特征样本。此外,根据当前位置内容所属标签(句末或非句末)对抽取出来的特征样本进行分类。若当前位置为句末,则下文信息由句首位置的对应内容代替,这样的样本称为正样本;若当前位置为非句末,由其组成的样本称为负样本。正负样本抽取完成后,平衡正负样本数量的比例,将数量接近的正负样本提供给模型进行学习。正负样本具体提取方式如式(3)所示:

$$Z_{x_i}=[x_{i-2},x_{i-1},x_i,x_{i+1},x_{i+2}] \tag{3}$$

其中,Z_{x_i} 是以 x_i 作为中心提取的样本。输入序列为 $[x_0,x_1,\cdots,x_n]$,窗口大小为 5,取当前位置前后两个位置的内容与当前位置组合在一起作为一组样本。

4)子词切分(Byte-Pair Encoding,BPE)。在对泰语句子进行句子编码之前,可以利用子词切分策略将泰语句子中的词进一步切分成精度更小的子词。BPE 方法经常被应用到深度学习模型的训练过程中,因为其能够将待训练的数据进一步细化成粒度更小的子词,缩小模型在训练过程中词汇表的规模,减少模型参数量,使模型在训练过程中能够更快地找到最优参数,加快模型收敛速度。

5)标签化。本文实验将泰语句子切分任务看作一个序列标注问题。因此,在训练开始前根据句子中的单位粒度进行标签映射,为每个单位位置赋予一个指定标签,使用 <B-Sen> 标志表示当前位置为一个句子的开头,使用 <O-Sen> 标志表示句子除了开头以外的其他位置,如图 4 所示。模型在训练完成后进行泰语句子切分时,可以利用模型输出的标签化序列,根据句子开头标志对当前处理的泰语句子序列进行切分。

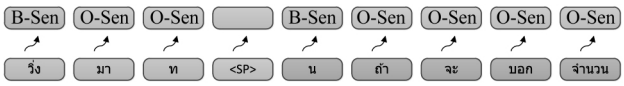

图 4 训练数据标签化

Fig. 4 Tagging of training data

4.2 实验环境与模型参数

本文实验采用如表 2 所示的系统环境以及如表 3 所示的超参数设置。

表 2 实验环境

Table 2 Experimental environment

系统环境	具体信息
操作系统	CentOS
GPU	NVIDIA GTX 1080Ti
Python	3.6.0
Tensorflow	1.12.0
内存/GB	256

表 3 模型超参数设置

Table 3 Model hyperparameter setting

超参数	参数值
batch_size	64
dropout	0.2
隐藏层维度	300
学习率	0.001

4.3 实验设计

本实验分为 3 组,具体设计如下:

1)使用不同方法构造模型网络结构,挑选出最优模型。本文实验对比了 CRF 模型、Bi-LSTM + Glove 模型和 Bi-LSTM + CRF + Glove 模型,分别记为 CRF、Bi-LSTM 和 Bi-LSTM + CRF。

2)对比不同粒度的句子切分方法在最优模型上的表现,选出最优的句子切分粒度。分别以基于词、字符以及子词的粒度对句子进行编码,记为 base-Word、base-Char 和 base-Bpe。

3)使用不同维度大小的句子编码方式对泰语句子进行编码并应用在最优模型上,选出最优的编码方式。实验使用的维度大小分别为 50、100、150、200、250 和 300,记为 Glove-50d、Glove-100d、Glove-150d、Glove-200d、Glove-250d 和 Glove-300d。

4.4 模型效果验证指标

实验使用准确率(precision)、召回率(recall)和 F1 值来衡量泰语分句模型的性能,F1 值为模型准确率和召回率的调和平均值,具体公式如下所示:

$$\text{precision}=\frac{T_p}{T_p+F_p}\times 100\% \tag{4}$$

$$\text{recall}=\frac{T_p}{T_p+F_n}\times 100\% \tag{5}$$

$$\text{F1}=2\times\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}\times 100\% \tag{6}$$

precision 表示模型预测的所有正样本边界中正确预测的边界数量。recall 表示所有真实正样本边界中被正确预测的边界数量。T_p 为模型识别正确的句子边界个数,F_p 为模型识别错误的句子边界个数,F_n 为是句子边界但模型未正确识别的个数。在分类问题中,F1 值能够综合准确率及召回率的结果,从而更加准确地反映出模型精度。

4.5 结果分析

分别使用 CRF、Bi-LSTM 和 Bi-LSTM + CRF 3 种模型结构完成泰语句子切分任务,效果对比如表 4 所示。可以看出,实验中的最优模型 Bi-LSTM + CRF 在泰语句子切分任务上能够达到非常精准的识别效果,明显优于另两种方法。

在最优模型上,对比不同粒度的句子切分方法,实验结果如表 5 所示。可以看出,基于词的句子编码形式能够获得最好的效果,这是因为以词为最小

切分单元进行编码并组成句子向量能够帮助模型在学习过程中获得更多词意信息,从而提升分句准确率。

表 4　3 种模型句子切分效果对比

Table 4　Comparison of sentence segmentation effect of three models　%

模型	准确率	召回率	F1 值
CRF	95.80	73.10	80.30
Bi-LSTM	96.00	82.90	88.90
Bi-LSTM + CRF	98.20	98.30	98.20

表 5　不同粒度的句子切分方法对模型的影响

Table 5　Influence of sentence segmentation methods with different granularity on the model　%

句子切分粒度	准确率	召回率	F1 值
base-Word	98.20	98.30	98.20
base-Char	96.48	92.30	94.34
base-Bpe	94.12	66.67	50.00

将泰语句子中的词转换成不同维度大小的编码向量供模型进行训练,探究单词编码维度的大小对模型效果的影响,实验结果如表 6 所示。实验结果表明,使用更高维度的编码方式由于将数据编码成高维向量,因此能够使模型得到更多的有效信息,从而更好地表示词意和句意信息,最终获得更高的切分精度。

表 6　不同句子编码维度下本文模型的表现

Table 6　Representation of the model in different sentence coding dimentions　%

句子编码维度	准确率	召回率	F1 值
Glove-50d	87.01	82.96	84.94
Glove-100d	88.69	82.73	85.60
Glove-150d	87.40	84.47	85.91
Glove-200d	97.98	98.36	98.17
Glove-250d	97.98	98.32	98.15
Glove-300d	98.27	98.30	98.28

5　结束语

泰语特有的语言特性使其句子切分任务具有复杂性。为此,本文使用 LSTM 神经网络的变形结构 Bi-LSTM + CRF 模型对泰语句子进行切分。实验表明,该模型对泰语能够进行准确切分。在此基础上,本文研究了使用不同粒度的句子切分方法和不同维度大小的编码方式在同一模型上的表现,证明基于词的句子切分方法和更高维度的向量能够使模型得到最高的准确率。下一步将针对与泰语具有相似语言特性的语种展开研究,例如高棉语、老挝语等,使得同一模型能够对多个语种的段落进行精确切分,提升模型的泛化性。

参考文献

[1] HE Qianhua, XU Bingzheng. Overview of machine translation [J]. Information Science, 1993, 11(4): 60-67. (in Chinese)
贺前华,徐秉铮. 机器翻译综述 [J]. 情报科学, 1993, 11(4): 60-67.

[2] GAO Minghu, YU Zhiqiang. A summary review of neural machine translation [J]. Journal of Yunnan University of Nationalities (Natural Sciences Edition), 2019, 28(1): 72-76. (in Chinese)
高明虎,于志强. 神经机器翻译综述 [J]. 云南民族大学学报: 自然科学版, 2019, 28(1): 72-76.

[3] PENG Shuchu. A review of the development of machine translation [J]. Journal of Huazhong University of Science and Technology (Social Science Edition), 2006, 20(2): 123-124. (in Chinese)
彭述初. 机器翻译学科发展综述 [J]. 华中科技大学学报(社会科学版), 2006, 20(2): 123-124.

[4] WANG R, UTIYAMA M, FINCH A, et al. Sentence selection and weighting for neural machine translation domain adaptation [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1727-1741.

[5] GIATSOGLOU M, VOZALIS M G, DIAMANTARAS K I, et al. Sentiment analysis leveraging emotions and word embeddings [J]. Expert Systems with Applications, 2017, 69: 214-224.

[6] ZHENG Lijuan, WANG Hongwei, GAO Song. Sentimental feature selection for sentiment analysis of Chinese online reviews [J]. International Journal of Machine Learning and Cybernetics, 2018, 9(1): 75-84.

[7] CHEN Tianying, CHEN Rong, PAN Lulu, et al. Archaic Chinese punctuating sentences based on context N-gram model [J]. Computer Engineering, 2007, 33(3): 192-193. (in Chinese)
陈天莹,陈蓉,潘璐璐,等. 基于前后文 n-gram 模型的古汉语句子切分 [J]. 计算机工程, 2007, 33(3): 192-193.

[8] XUE Zhengshan, ZHANG Dakun, WANG Lina, et al. An improved sentence segmentation model for machine translation [J]. Journal of Chinese Information Processing, 2017, 31(4): 50-56. (in Chinese)
薛征山,张大鲲,王丽娜,等. 改进机器翻译中的句子切分模型 [J]. 中文信息学报, 2017, 31(4): 50-56.

[9] WIROTE A. Thoughts on word and sentence segmentation in Thai [C]//Proceedings of the 7th International Symposium on Natural Language Processing. Pattaya, Thailand: [s. n.], 2007: 85-90.

[10] KASISOPA B, REILLY R, BURNHAM D. Orthographic factors in reading Thai: an eye tracking study [C]// Proceedings of the 4th China International Conference on Eye Movements. Tianjin, China: [s. n.], 2010: 1-2.

[11] HALTERENH V. Syntactic wordclass tagging [M]. 1st editon. Amsterdam, Holland: Kluwer Academic Publishers, 1999.

[12] SILLA C N, KAESTNER C A. An analysis of sentence boundary detection systems for English and Portuguese documents [C]//Proceedings of the 5th International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Germany: Springer, 2004: 135-141.

[13] JURAFSKY D, MARTIN J H. Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition [M]. 2nd edition. Upper Saddle River, USA: Prentice Hall, 2008.

[14] CHAROENPORNSAWAT P, SORNLERTLAMVANICH V. Automatic sentence break disambiguation for Thai [EB/OL]. [2019-07-10]. http://www.cs.cmu.edu/~paisarn/papers/iccpol2001.pdf.

[15] MITTRAPIYANURUK P, SORNLERTLAMVANICHV. The automatic Thai sentence extraction [C]//Proceedings of the 4th Symposium on Natural Language Processing. Chiang Mai, Thailand: [s. n.], 2000: 1-6.

[16] SLAYDEN G, HWANG M Y, SCHWARTZ L. Thai sentence-breaking for large-scale SMT [C]//Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing. [S. l.]: The COLING 2010 Organizing Committee, 2010: 8-16.

[17] WATHABUNDITKUL S. Spacing in the Thai language [EB/OL]. [2019-07-10]. http://www.thai-language.com/ref/spacing.

[18] ZHOU N, AW A, LERTCHEVA N, et al. A word labeling approach to Thai sentence boundary detection and POS tagging [C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee, 2016: 319-327.

[19] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [EB/OL]. [2019-07-10]. http://www.cs.cmu.edu/~xuezhem/publications/P16-1101.pdf.

[20] HAN Xuedong. Summary of conditional random field theory [EB/OL]. [2019-07-10]. https://wenku.baidu.com/view/842401c42cc58bd63186bd4b.html. (in Chinese)
韩雪冬. 条件随机场理论综述 [EB/OL]. [2019-07-10]. https://wenku.baidu.com/view/842401c42cc58bd63186bd4b.html.

[21] ZHANG Zirui, LIU Yunqing. Chinese word segmentation based on bi-directional LSTM-CRF model [J]. Journal of Changchun University of Science and Technology (Natural Science Edition), 2017, 40(4): 87-92. (in Chinese)
张子睿, 刘云清. 基于 BI-LSTM-CRF 模型的中文分词法 [J]. 长春理工大学学报(自然科学版), 2017, 40(4): 87-92.

编辑 任欣平

(上接第 293 页)

[14] JOSHI M, HADI T. A review of network traffic analysis and prediction techniques [EB/OL]. [2019-06-25]. http://www.oalib.com/paper/4048815#.X0SSgFN_n3Q.

[15] LI Peiyu. Short-term traffic flow prediction based on wavelet and neural network [J]. Computer Technology and Development, 2020, 30(1): 135-139. (in Chinese)
李佩钰. 一种基于小波和神经网络的短时交通流量预测 [J]. 计算机技术与发展, 2020, 30(1): 135-139.

[16] WANG Pengtao, HAN Xiaoming, HE Min. Prediction simulation of bicycle demand for public bicycle stations [J]. Computer Simulation, 2019, 36(8): 421-426, 458. (in Chinese)
王鹏涛, 韩晓明, 贺敏. 公共自行车站点需求量预测仿真 [J]. 计算机仿真, 2019, 36(8): 421-426, 458.

[17] JIA Hengjian. Investigation into the effectiveness of long short term memory networks for stock price prediction [EB/OL]. [2019-06-25]. https://arxiv.org/abs/1603.07893v1.

[18] ZHAO Zheng, CHEN Weihai, WU Xingming, et al. LSTM network: a deep learning approach for short-term traffic forecast [J]. IET Intelligent Transport Systems, 2017, 11(2): 68-75.

[19] KOZIK R. Distributing extreme learning machines with apache spark for NetFlow-based malware activity detection [J]. Pattern Recognition Letters, 2018, 101(1): 14-20.

[20] ZHANG Junbo, ZHENG Yu, QI Dekang. Deep spatio-temporal residual networks for citywide crowd flows prediction [EB/OL]. [2019-06-25]. https://www.researchgate.net/publication/308809186_Deep_Spatio-Temporal_Residual_Networks_for_Citywide_Crowd_Flows_Prediction.

[21] SHAO H X, SOONG B H. Traffic flow prediction with Long Short-Term Memory Networks (LSTMs) [C]//Proceedings of TENCON'16, Washington D. C., USA: IEEE Press, 2016: 2986-2989.

[22] TULI H, KUMAR S. Prediction analysis of delay in transferring the packets in adhoc networks [C]//Proceedings of 2016 International Conference on Computing for Sustainable Global Development. Washington D. C., USA: IEEE Press, 2016: 35-42.

[23] FENG Ning, GUO Shengnan, SONG Chao, et al. Multi-component spatial-temporal graph convolution networks for traffic flow forecasting [J]. Journal of Software, 2019, 30(3): 759-769. (in Chinese)
冯宁, 郭晟楠, 宋超, 等. 面向交通流量预测的多组件时空图卷积网络 [J]. 软件学报, 2019, 30(3): 759-769.

编辑 宋 圆