基于 BLSTM-ATT 的老挝语军事领域实体关系抽取

何阳宇[1],易晓宇[1],唐 亮[1],易绵竹[1],李宏欣[1,2]
(1. 解放军战略支援部队信息工程大学 洛阳校区,河南 洛阳 471003;
2. 密码科学技术国家重点实验室,北京 100878)

摘 要:为了对互联网上大量的老挝语军事类文本进行结构化分析,该文提出了一种基于双向长短期记忆网络和多头自注意力机制的军事领域实体关系抽取方法。针对老挝语语料匮乏问题,提出了“硬匹配”和“软匹配”的思想,在完成语料获取和预处理的基础上,利用预定义的关系词表进行“硬匹配”,之后再通过词典匹配和相似度计算相结合的方法进行“软匹配”,以提高关系类型的泛化能力,进而自行构建了关系抽取标注语料库;然后,通过分析老挝语语言特点,融入了词、词性、实体类型、相对位置关系等特征进行模型训练,并设置了四轮针对不同变量的对比实验,验证了不同的神经网络模型、注意力机制、嵌入的特征以及语料规模对抽取效果的影响程度,实验结果表明融合双向长短期记忆网络和多头自注意力的方法对老挝语军事领域实体关系抽取具有更好的性能。

关键词:双向长短期记忆网络;多头注意力;老挝语;军事领域;实体关系抽取
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2021)05-0031-07
doi:10.3969/j.issn.1673-629X.2021.05.006

Lao Entity Relation Extraction in Military Domain Based on BLSTM and Attention Mechanism

HE Yang-yu[1], YI Xiao-yu[1], TANG Liang[1], YI Mian-zhu[1], LI Hong-xin[1,2]
(1. Luoyang Campus, PLA Strategic Support Force Information Engineering University, Luoyang 471003, China;
2. State Key Laboratory of Cryptology, Beijing 100878, China)

Abstract: In order to conduct a structured analysis of a large number of Lao military texts on the Internet, we propose a method for extracting entity relationship in military domain based on a bidirectional LSTM network and multi-head self-attention mechanism. In view of the lack of Lao corpus, the idea of “hard matching” and “soft matching” is proposed. Based on the completion of corpus acquisition and preprocessing, the pre-defined relation vocabulary is used for “hard matching”, and then the method combined dictionary matching with similarity calculation is used to perform “soft matching” to improve the generalization ability of relation types, then a relation extraction corpus is built on its own. By analyzing the characteristics of Lao language, the features such as word, part of speech, entity type and relative position relationship are incorporated for model training, and four rounds of comparative experiments for different variables are set up to verify the effect of different neural network models, attention mechanisms, embedded features and corpus size on the extraction effect. The experiment shows that the fusion of the bidirectional LSTM network and multi-head self-attention mechanism have better performance for Lao entity relation extraction in military domain.

Key words: bidirectional long short-term memory networks; multi-head attention; Lao; military domain; entity relation extraction

0 引 言

随着信息时代的快速发展,互联网上与军事相关的内容大量涌现,其中以老挝语形式发布的信息也越来越多,同时带来了信息过载、不直观、利用率低等问题,人工很难全面、准确、及时地对要点进行处理分析。因此,自动地从非结构化文本中抽取所需信息成为解决这一难题的关键。

实体关系抽取作为信息抽取的主要任务之一,其

收稿日期:2020-03-18 **修回日期**:2020-07-19
基金项目:国家自然科学基金项目(61701539);国防科技创新特区项目(18-H863-01-ZT-005-005-01)
作者简介:何阳宇(1992-),男,博士研究生,CCF 会员(B7818G),研究方向为自然语言处理和知识图谱;易绵竹,博士,教授,研究方向为语义学和计算语言学。

目的是从非结构化文本中抽取实体之间显式或隐式的语义关联,解决关系分类问题。作为信息抽取的关键环节,关系抽取在语义检索、自动问答、知识图谱等诸多领域具有广阔的应用前景。尤其对于老挝语来说,大量的文本尚以非结构化的形式存在,迫切需要进行深入挖掘分析,为促进老挝语自然语言处理研究以及知识库构建提供数据基础和技术支撑。

关系抽取经历了基于规则模板的方法、基于传统统计模型的方法,发展到如今的深度学习方法,在英语、汉语等大语种中已经取得了很好的效果。基于深度学习的关系抽取又可分为有监督、弱监督和无监督三种,其中有监督的方法目前具有更高的准确率和召回率,但是需要大量的标注语料。老挝语作为小语种,研究基础薄弱,标注语料匮乏,为解决这一问题,该文提出了“硬匹配(hard matching)”与“软匹配(soft matching)”相结合的启发式方法,自行构建了用于关系抽取的标注语料库。

在自建语料库的基础上,该文提出将双向长短期记忆网络模型 BLSTM 和注意力机制用于老挝语军事领域实体关系的抽取,BLSTM 可以从正反两个方向学习上下文特征,较好地捕捉双向的长距离语义依赖关系,非常适合长句数量众多的老挝语文本,而采用的多头注意力机制能够允许模型在不同的位置关注来自不同表征子空间的信息,进一步突出了老挝语句子中对语义关系具有重要影响的信息,克服了单头注意力机制只能取平均值,从而可能导致某些重要信息被掩盖的缺陷。

1 相关研究

关系抽取的方法主要可分为两大类:基于规则的方法和基于机器学习的方法。早期的关系抽取大多利用规则在文本中寻找与其相匹配的实例,从而推导出实体之间的语义关系。文献[1]根据实体之间的谓语动词来判断它们的关系,文献[2]通过语义注释句法树生成规则进行实体关系识别,此类方法对于特定领域的关系抽取准确率较高,但其扩展性和移植性较差,召回率普遍较低。文献[3]综合实体本身、实体类型、依存树和解析树等特征建立最大熵模型来判断实体关系类型,文献[4]在文本浅层解析的基础上定义树核函数,并结合支持向量机和投票感知器抽取实体关系。随着深度学习的兴起,基于神经网络的关系抽取成为近年来的研究热点,文献[5]提出利用卷积神经网络(CNN)进行关系抽取,将词汇级别特征和句子级别特征拼接得到的向量输入 softmax 分类器中预测实体关系,文献[6]引入循环神经网络(RNN)为解析树的每个节点分配向量和矩阵,并通过模型学习命题逻辑和自然语言中运算符的含义,以此产生任意句法类型和长度的短语和句子的组合向量表示,最后用 softmax 进行关系分类。理论上,RNN 可以处理任意长度的序列数据,但在实际操作中,当有用信息距离当前处理信息较远时,就容易导致 RNN 产生梯度消失或梯度爆炸等问题。针对这一现象,文献[7]提出长短期记忆网络(LSTM),通过引入门控机制大幅度提高了处理长序列数据的能力。此后,有研究者将 LSTM 改进为 BLSTM,并将其用于关系分类,如文献[8]利用双向长短期记忆网络(BLSTM)对完整序列进行建模,并嵌入了词汇、句法等特征,实验结果表明该方法进一步提升了关系分类性能。近年来,注意力机制也成功应用到关系分类任务中,文献[9]将 BLSTM 与注意力机制结合起来进行关系抽取,注意力层可以对 BLSTM 网络的输出进行加权变换,获取句子中每个词对语义关系的影响力权重,从而获得更加准确的分类结果。该文在其基础上,使用多头自注意力机制,能够更多地关注序列内部结构,各注意力头不仅能执行不同的任务,还能在一定程度上体现句法和语义特征。

另一方面,目前暂无老挝语关系抽取方面的研究,但是与关系抽取任务密切相关的分词、词性标注和命名实体识别等已有相关成果,这为该文提供了较好的基础条件。此外,作为同源语言的泰语在关系抽取方面有少量研究可供参考,文献[10]利用基于特征的方法在与犯罪相关的新闻语料中进行关系抽取,文献[11]提出基于最大熵的泰语句子级实体从属关系抽取方法,以汉泰平行句对作为桥梁构建语料库,然后选择符合泰语特点的上下文特征,使用最大熵模型进行训练,取得了不错的效果,该方法对该文有一定的启示意义,但缺乏高质量的汉老平行语料,因此不能完全适用。

综上,该文构建了一种基于 BLSTM 和多头自注意力机制的老挝语军事领域实体关系抽取模型,首先采用半自动的方法自行构建了相关语料库,在一定程度上解决了标注语料稀缺的问题,然后利用模型进行训练,最后输入 softmax 分类器进行关系分类。考虑到老挝语的研究现状和应用需求,该文主要探讨句子级的二元关系抽取。

2 老挝语军事领域实体关系语料库构建

英语等大语种已有 SemEval - 2010 Task 8、ACE2004 等专门用于关系抽取研究的公开标注数据集,而老挝语暂无类似资源。因此,该文需自行构建相关语料库——LREC(Lao relation extraction corpus),流程见图 1。

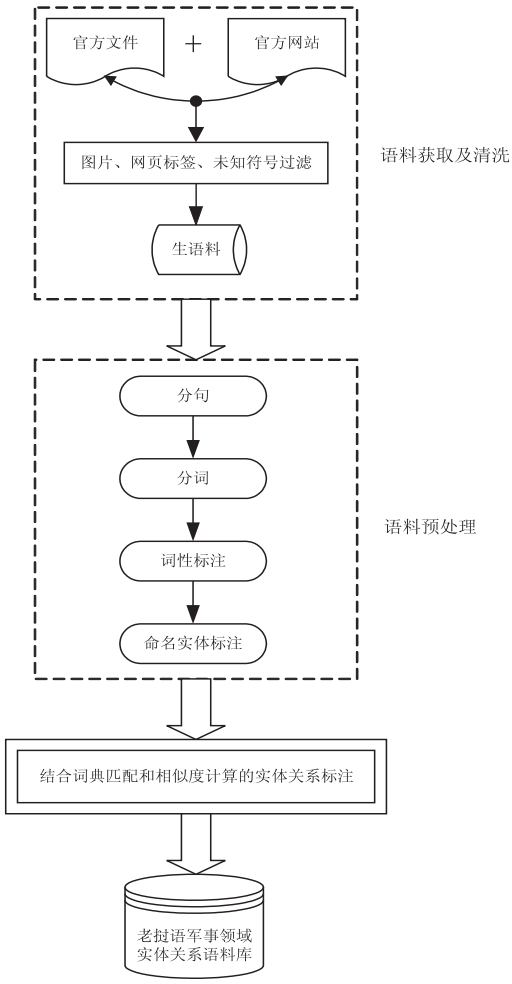

图 1　老挝语军事领域实体关系语料库构建流程

2.1　语料获取及清洗

语料的主要来源为老挝国防部、人民军、老挝通讯社等官方网站的军事类新闻以及《老挝国防政策白皮书》、《国防法》等官方文件,这些语料具有相关性强、质量高等特点,符合该文需求。确定语料来源后,利用爬虫技术进行爬取,然后过滤掉图片、网页标签、未知符号等噪声,最后得到所需的生语料。

2.2　语料预处理

得到生语料后需对其进行一系列的预处理:第一步是分句,以句号、问号、感叹号等作为标志对文本进行切分,得到约 11 500 个句子组成的句子集。第二和第三步分别是对句子集进行分词和词性标注,现有的相关工具(由昆明理工大学信息工程与自动化学院开发。)主要面向通用领域,为提高在军事领域的性能,该文在其基础上融入了领域词表,包括《老汉-汉老军事词典》[12]以及自建的老挝军事领域人名、地名库等。第四步是命名实体标注,方法沿用文献[13]。最后对语料库作进一步筛选,将所含实体数少于 2 个的句子剔除,操作后剩下 9 211 个句子。

2.3　结合词典匹配和相似度计算的实体关系标注

首先,根据老挝语的语言特点和军事领域的任务需求,预定义了 10 种关系类型,见表 1。

首先根据表 1 构建表达关系类型的关系词词表,然后在已完成预处理的语料中执行“硬匹配”操作,即精确的字符串匹配,接着输出关系实例,形式化表达为 $(E_1,R,E_2;S)$,其中 E_1 和 E_2 为两个实体,R 为关系,S 代表包含 E_1 和 E_2 的句子。由于语言表达的多样性,如果只进行硬匹配,会降低每种关系类型的泛化能力,一些语义相似但文字表征不同的关系词可能会被忽略,从而影响关系标注语料库的质量和规模。因此,还需要进行“软匹配”。“软匹配”分为两部分,即基于词典的匹配和基于 Word2vec 的相似词推荐。基于词典的匹配原理是利用词表中关系词的汉语释义在《老挝语汉语词典》[14]中查找同义词或近义词。比如,“创建”的老挝语表达有“ສ້າງຕັ້ງ”、“ສະຖາປະນາ”等,查找完毕后都归入“ສ້າງຕັ້ງ”这一关系类型;基于 Word2vec 的相似词推荐是利用以上获得的关系词及其同义词和近义词作为种子词集进行相似词推荐,为保证质量,将推荐阈值设定为 5,之后也将推荐词归入相应的关系类型。最后,将“软匹配”获得的关系词同样带入语料库中,输出关系实例。经过以上步骤,对未包含表 1 中任何一种关系类型的句子再次执行过滤操作。最终,经老挝语专家人工抽样检查后得到可用于关系抽取的标注语料 5 063 句。

表 1　老挝语军事领域实体关系类型

关系类型	可连接实体类型示例
ຮ່ວມມື(合作关系)	军事机构-军事机构
ຂຶ້ນກັບ(隶属关系)	军事机构-军事机构
ຂັ້ນເທິງ-ຂັ້ນລຸ່ມ(上下级关系)	人物-人物
ສ້າງຕັ້ງ(创建关系)	人物-军事机构
ຜະລິດ(生产制造关系)	军事机构-武器装备
ພັດທະນາ(开发关系)	军事机构-武器装备
ປະກອບ(配备关系)	人物-武器装备
ນຳໃຊ້(使用关系)	人物-武器装备
ຈັດວາງ(部署关系)	军事设施-地点
ຕັ້ງຢູ່(位于关系)	军事机构-地点

3　基于 BLSTM-ATT 的老挝语军事领域实体关系抽取模型

基于 BLSTM 和多头注意力机制的老挝语军事领域实体关系抽取模型整体架构见图 2。该模型的第一步是将经过预处理的句子输入模型,提取初始特征;第二步通过嵌入层(embedding layer)将所有初始特征映射为低维稠密向量;第三步利用 BLSTM 从第二步获取高层特征;第四步是引入注意力机制产生权重向量,并将其与 BLSTM 层输出的向量加权求和,形成更高层次的特征向量;最后将所得向量输入 softmax 分类器用于关系分类。

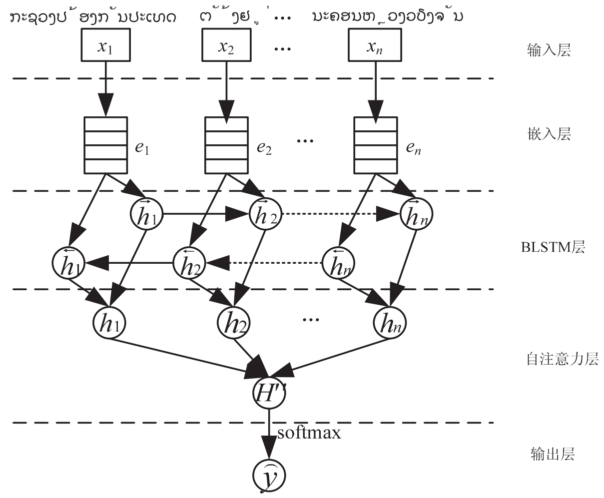

图 2 基于 BLSTM-ATT 的老挝语军事领域实体关系抽取模型架构

3.1 初始特征提取

常用于关系抽取的特征包括词、词性、实体类型、上下位关系、相对位置关系、依存关系和语义角色等。由于老挝语缺乏依存解析、语义角色标注等工具,该文选择词、词性、实体类型、相对位置关系作为初始特征,前三项特征可利用已有的自然语言处理工具得到,而位置特征则通过计算当前词 x_t 到实体 E_1 和 E_2 的相对距离获取。给定老挝语句子 $S = \{x_1, x_2, \cdots, x_n\}$,其中 n 表示 S 中包含的词数,那么 x_t 到 E_1 和 E_2 的相对距离 D_1、D_2 可分别通过式(1)、式(2)得出,式中 t_1 和 t_2 分别是 E_1 和 E_2 的索引下标,结果为负代表 x_t 位于实体前,反之则位于实体后,见图 3。位置特征的加入使模型更加明确了哪两个实体词需要进行关系分类。

$$D_1 = t - t_1 \tag{1}$$

$$D_2 = t - t_2 \tag{2}$$

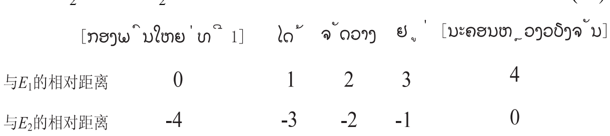

图 3 位置特征示例

综上,从 S 中提取到的初始特征集可表示为 $K = \{k_1, k_2, \cdots, k_q\}$,其中 q 为特征集大小,取值为 4。

3.2 特征嵌入

特征嵌入就是将初始特征映射为实数向量。通过训练模型可以将 S 转化为一个可学习的多维参数矩阵 $\boldsymbol{W} \in R^{d \times |V|}$,其中 d 表示词向量维度, V 表示词表大小, x_t 通过矩阵向量积运算便可得到对应的词向量表示 r_t ,如式(3),其中 v_t 是关于 x_t 的独热表示。

$$r_t = \boldsymbol{W} v_t \tag{3}$$

同理,可以得出所有初始特征的向量表示 $r_t^{k_j}$ 的计算方法,如式(4),其中 k_j 表示第 j 种初始特征类型, $W^{k_j} \in R^{d^{k_j} \times |V^{k_j}|}$ 、d^{k_j} 和 V^{k_j} 分别表示相应特征的向量维度和取值类别数。

$$r_t^{k_j} = W^{k_j} v_t^{k_j} \tag{4}$$

将所有特征向量拼接起来便可以得到 x_t 完整的向量化表示 $e_t = \{r_t^{k_1}, r_t^{k_2}, \cdots, r_t^{k_q}\}$,最终 S 的向量表示为 $e_s = \{e_1, e_2, \cdots, e_n\}$ 。

3.3 BLSTM 层

LSTM 本质上是一种 RNN,只是在标准 RNN 的基础上引入了门控机制,包括遗忘门(forget gate)、输入门(input gate)和输出门(output gate),这种“门结构”相当于一种过滤装置,能够保留重要信息,丢弃不重要信息,其具体结构见图 4。一个 LSTM 模块在每个时间步会接收三个输入,即当前时间步的输入 e_t 、上一个时间步的内部状态 c_{t-1} 以及上一个时间步的外部状态 h_{t-1} ,输出则包括当前时间步的内部状态 c_t 和外部状态 h_t 。另外, f 、i 、$\tilde{c}_t$ 、o 分别表示遗忘门、输入门、备选状态、输出门。

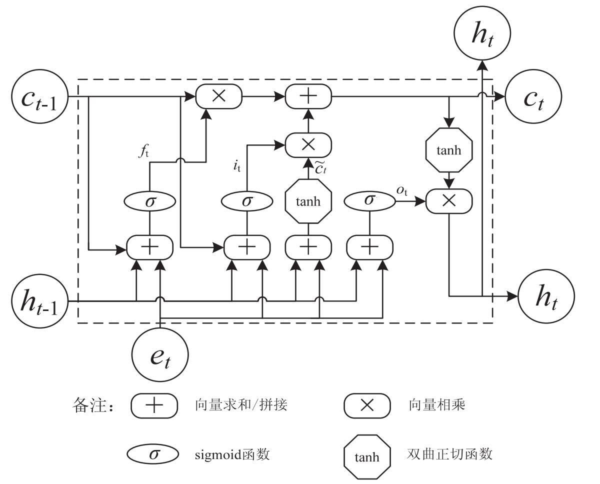

图 4 LSTM 模块结构

将嵌入层得到的向量表示 e_t 作为 LSTM 层的输入,计算第 t 个词时 LSTM 各个状态特征值的过程如下所示:

第一步由遗忘门决定上一个时间步内部状态信息的去向,计算方法如式(5), $\boldsymbol{W}_{xf}$ 、$\boldsymbol{W}_{hf}$ 、$\boldsymbol{W}_{cf}$ 和 b_f 为 f_t 对应的权重矩阵和偏置(bias),式(6) ~ 式(9)中的类似符号不再赘述。

$$f_t = \sigma(\boldsymbol{W}_{xf} e_t + \boldsymbol{W}_{hf} h_{t-1} + \boldsymbol{W}_{cf} c_{t-1} + b_f) \tag{5}$$

第二步决定更新哪些信息,包括两层操作:一是由输入门决定需要更新的信息,如式(6);二是通过 tanh 层产生用于更新的 $\tilde{c}_t$,如式(7)。这样就完成了对内部状态进行更新的准备。

$$i_t = \sigma(\boldsymbol{W}_{xi} e_t + \boldsymbol{W}_{hi} h_{t-1} + \boldsymbol{W}_{ci} c_{t-1} + b_i) \tag{6}$$

$$\tilde{c}_t = \tanh(\boldsymbol{W}_{xc} e_t + \boldsymbol{W}_{hc} h_{t-1} + \boldsymbol{W}_{cc} c_{t-1} + b_c) \tag{7}$$

第三步是对内部状态进行更新,即将 c_{t-1} 更新为 c_t ,如式(8)。

$$c_t = i_t \tilde{c}_t + f_t c_{t-1} \tag{8}$$

第四步是输出。首先,确定内部状态的哪个部分被输出,如式(9),然后利用 tanh 层对当前时间步的内部状态 c_t 进行处理得到最终输出,如式(10)。

$$o_t = \sigma(\boldsymbol{W}_{xo} e_t + \boldsymbol{W}_{ho} h_{t-1} + \boldsymbol{W}_{co} c_t + b_o) \tag{9}$$

$$h_t = o_t \tanh(c_t) \tag{10}$$

由于时序信息是按照时间从前往后依次传播,LSTM 只能依据之前时刻的信息来预测当前时刻的输出,但在序列建模任务中,当前时刻的输出往往不仅与之前的状态有关,还与未来的状态有关。BLSTM 正是为解决这一问题提出的,它由一个正向 LSTM 和一个反向 LSTM 组成,这两个独立的循环网络分别负责学习上文和下文的特征信息,最后拼接起来送入同一输出层。在式(10)的基础上可以得出,利用 BLSTM 处理句子 S 时,第 t 个词的输出如式(11):

$$h'_t = [\overrightarrow{h}_t \oplus \overleftarrow{h}_t] \tag{11}$$

综上,句子 S 的输出向量为 $\boldsymbol{H} = \{h'_1, h'_2, \cdots, h'_n\}$。

3.4　自注意力层

老挝语文本中常出现“ທີ່”、“ຊຶ່ງ”、“ອັນ”等引导的从句,造成长句现象较多,而 BLSTM 对这些句子中的所有词都“一视同仁”,无法突出对语义关系具有重要影响的信息。因此,该文添加了多头自注意力层,对 BLSTM 的输出进行加权变换,获取句子中每个词对语义关系的影响力权重。

多头自注意力层处理 $\boldsymbol{H} = \{h'_1, h'_2, \cdots, h'_n\}$ 的过程如式(12)~式(16)所示:

$$M = \tanh(\boldsymbol{H}) \tag{12}$$

$$A = \mathrm{softmax}(\boldsymbol{w}^n M) \tag{13}$$

$$\boldsymbol{B} = \boldsymbol{H}\boldsymbol{A}^n \tag{14}$$

$$\boldsymbol{H}' = \tanh(\boldsymbol{B}) \tag{15}$$

$$\boldsymbol{H}'' = \mathrm{concat}(\boldsymbol{H}'_1, \boldsymbol{H}'_2, \cdots, \boldsymbol{H}'_l) \odot \boldsymbol{w}' \tag{16}$$

以上公式中,$\boldsymbol{A}$ 表示注意力权重矩阵,$\boldsymbol{w}$ 是训练所得的参数向量,$\boldsymbol{w}^n$ 是 $\boldsymbol{w}$ 的转置(transpose),$\boldsymbol{B}$ 表示完成加权变换后所得的句子向量,$\boldsymbol{H}'$ 表示单一注意力头得到的输出特征,假设一共进行 l 次注意力计算,concat 表示向量拼接,⊙表示逐元素点乘,最终得到的输出为 $\boldsymbol{H}''$。

3.5　输出层

老挝语关系抽取实际上是一个分类问题,输出层为 softmax 分类器,将注意力层得到的 $\boldsymbol{H}''$ 输入其中便得到每个关系类别的条件概率,取其中概率最大的作为模型最终预测结果。计算过程如式(17)~式(19)所示:

$$p(y \mid \boldsymbol{H}'') = \mathrm{softmax}(\boldsymbol{W}^s \boldsymbol{H}'') + b^s \tag{17}$$

通过式(17)可以计算出句子 S 属于各个关系类别的概率,其中 $\boldsymbol{W}^s \in R^{z \times ld^h}$,$z$ 是预定义的关系类别数量。然后通过式(18)得到概率最大的类别 $\hat{y}$。

$$\hat{y} = \underset{y}{\mathrm{argmax}}(y \mid \boldsymbol{H}'') \tag{18}$$

为了优化模型,采用带有 L2 惩罚项的交叉熵(cross entropy)损失作为目标函数,如式(19)所示:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda \|\theta\|_F^2 \tag{19}$$

其中,θ 为模型参数,m 为样本数,$t_i \in R^m$ 是正样例的独热向量表示,$y_i \in R^m$ 是 softmax 计算出的每个关系类别的概率,λ 是 L2 正则化超参数。

4　实　验

4.1　实验语料

实验所采用的语料均来自 LREC,共包含 5 063 条数据,10 种预定义的关系类别,按照 4∶1 的比例分配训练语料和测试语料。

4.2　评价指标

为了综合评价系统性能,将准确率(P)、召回率(R)以及 F 测度值(F-measure)作为评价指标对模型进行测试,具体定义分别如式(20)~式(22):

$$P = \frac{\text{预测正确的关系数量}}{\text{预测出的关系总量}} \times 100\% \tag{20}$$

$$R = \frac{\text{预测正确的关系数量}}{\text{测试语料中存在的关系总量}} \times 100\% \tag{21}$$

$$F = \frac{2PR}{P + R} \times 100\% \tag{22}$$

4.3　参数设置

选用批量的 Adam[15] 优化方法训练模型,其中带有交叉熵损失函数。由于实验的参数较多,而用于老挝语关系抽取的语料相对较少,容易产生过拟合现象,因此,采用 L2 正则化来限制模型参数值,并且在嵌入层、BLSTM 层和注意力层使用“丢弃法(dropout)”[16] 策略。另外,将修正线性单元(rectified linear unit,ReLU)作为激活函数。对于具体参数值的设置,利用 K 折交叉验证法和参考过往研究经验值的方法共同完成,见表 2。

表 2　参数取值情况

参数名称	取值
学习率	0.01
批量大小	50
隐藏层节点数	300
正则化系数	0.000 01
丢弃率	0.3
词嵌入维度	300
词性嵌入维度	25

续表 2

参数名称	取值
实体类型特征嵌入维度	5
位置特征嵌入维度	25
多头自注意力并行头数	8

4.4 实验结果与分析

为了全方位地验证该方法的有效性,一共设置四组对比实验,变量分别是神经网络类型(实验一)、注意力机制(实验二)、嵌入特征(实验三)和语料规模(实验四),其他实验环境和参数设置等客观因素均保持一致,CNN 等需特殊设置的方法另行阐述。

4.4.1 实验一

本轮实验选取了当前常见的几种神经网络模型:CNN、RNN、LSTM 和 BLSTM。其中 CNN 的架构以及滤波器窗口尺寸和卷积核个数等设置借鉴文献[5],RNN 方法借鉴文献[17],以上模型均暂不加入注意力机制,具体结果见表 3。

表 3 实验一结果对比

序号	方法	准确率 P/%	召回率 R/%	F 值/%
1	CNN	69.85	69.34	69.59
2	RNN	71.07	72.49	71.77
3	LSTM	74.25	74.41	74.33
4	BLSTM	79.08	81.70	80.37

从表 3 可知,从 1 到 4 号实验结果总体呈上升趋势。具体来看,几种模型中 CNN 的效果最差,这是因为老挝语中长句较多,而 CNN 只能处理其窗口内的信息,难以应对长程依赖问题。RNN 比 CNN 的结果略有上升,证明 RNN 可以在一定程度上缓解较长序列建模问题,但是提升能力有限,这是由于当老挝语句子过长时,RNN 会出现梯度消失或梯度爆炸等问题。加入门控机制后的 LSTM,其结果有较大幅度的提高,这说明 LSTM 更适合处理时序数据。BLSTM 方法的各项指标在单向 LSTM 的基础上又有了 6% 左右的提升,这意味着在老挝语句子中两个方向的语义信息对抽取结果都非常重要,尤其是在句子结构较长且复杂的情况下,BLSTM 在充分利用上下文信息方面更具优势。

4.4.2 实验二

为了验证注意力机制的有效性以及注意力头数的作用,本轮实验设计了三个模型,分别是 BLSTM、BLSTM+ATT 和 BLSTM+Multi-Head ATT,具体结果见表 4。

从表 4 可看出,加入注意力后,抽取结果指标均提升了 4% 左右,这说明注意力机制能够充分获取到老挝语句子内部的有用特征,在较大程度上排除冗余信息的干扰。将单一注意力扩展为多头注意力后,效果更加显著,证明了多头注意力机制可以更加全面地捕捉句子信息,进一步提高模型的特征表达能力。

表 4 实验二结果对比

序号	方法	准确率 P/%	召回率 R/%	F 值/%
1	BLSTM	79.08	81.70	80.37
2	BLSTM+ATT	83.47	85.22	84.34
3	BLSTM+Multi-Head ATT	85.39	86.01	85.70

4.4.3 实验三

将词、词性、实体类型、相对位置关系四个特征分别记为 k_1、k_2、k_3、k_4。为了测试各特征对结果的影响程度,选用“BLSTM+Multi-Head ATT”作为训练模型,依次加入以上特征进行对比。具体结果见表 5。

表 5 实验三结果对比

序号	融入特征	准确率 P/%	召回率 R/%	F 值/%
1	k_1	77.51	77.88	77.69
2	$k_1 + k_2$	80.49	81.04	80.76
3	$k_1 + k_2 + k_3$	84.21	85.80	85.00
4	$k_1 + k_2 + k_3 + k_4$	85.39	86.01	85.70

从表 5 整体情况看,所选特征都是有效的,其中,词性特征 k_2 和实体类型特征 k_3 的加入对性能的提升较为明显,这是因为老挝语中表达语义关系的词一般是动词,k_2 可以帮助系统捕捉词性信息,而在军事领域关系抽取任务中,实体类型信息对正确预测关系也有较为显著的作用,比如,“部署”关系一般连接的是“武器装备”类实体和“地点”类实体。相比之下,位置特征 k_4 的加入对效果的提升帮助不大,这可能是因为老挝语中部分词的位置复杂多变,在一定程度上增加了 k_4 的不确定性。

4.4.4 实验四

为了考察语料规模对系统的影响,本轮实验以 500 条训练语料为起点,以 500 条为单位增量进行模型训练,选用“BLSTM+Multi-Head ATT”作为训练模型,共有约 4 000 条训练语料,测试语料规模保持不变,其余参数设置和实验环境等均一致,结果对比见图 5。

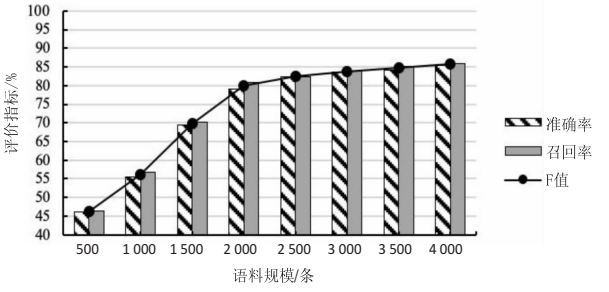

图 5 实验四结果对比

从图 5 可以看出,总体上的趋势是随着语料规模的增大,各项评价指标也随之提高,这说明训练语料的数量是提高老挝语军事领域关系抽取系统性能的关键因素。另一方面,在语料规模约为 2 000 条的时候, F 值就达到了 80% 左右,这说明“BLSTM + Multi - Head ATT”的方法能够在语料规模较小的时候,比较全面深入地挖掘已有语料的上下文信息和内部特征,实现较好的效果,非常适合老挝语语料匮乏的研究现状,但是在达到 2 000 条之后,评价指标的增速开始放缓,这可能是因为已有语料的句法类型等元素的丰富程度还不够,模型能够学习到的特征已经达到相对饱和的状态。

5　结束语

重点研究探讨了老挝语军事领域实体关系抽取问题。针对语料匮乏的情况,利用半自动的方法自行构建了关系抽取语料库,然后提出了基于 BLSTM 和多头自注意力的老挝语军事领域实体关系抽取模型,并结合老挝语语言特点和研究现状引入了词、词性、实体类型、相对位置关系等特征,最后进行了四轮对比实验,其结果证明了该模型的有效性和可靠性。尽管如此,仍然有较大的改进空间。比如,除了提及的特征以外,囿于基础研究薄弱,还有依存分析、语义角色等特征未能使用,待今后相关工具完善后可融入其中。作为低资源语言的老挝语,语料问题也一直是困扰研究的难点,今后除了继续加大语料建设以外,还要积极探索更多对语料依赖较小的方法。另外,将研究从军事领域扩展到其他领域甚至开放领域也是将来的研究方向。

参考文献:

[1] FUKUMOTO J, MASUI F, SHIMOHATA M, et al. Oki electric industry: description of the Oki system as used for MUC-7 [C]//Proceedings of the 7th message understanding conference. Fairfax, Virginia: [s. n.], 1998: 1-7.

[2] MILLER S, FOX H, RAMSHAW L, et al. A novel use of statistical parsing to extract information from text[C]//Proceedings of the 1st North American chapter of the association for computational linguistics conference. Seattle, Washington: ACL, 2000: 226-233.

[3] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations [C]//Proceedings of the ACL 2004 on interactive poster and demonstration sessions. Barcelona, Spain: ACL, 2004: 22.

[4] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003, 3(3): 1083-1106.

[5] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th international conference on computational linguistics. Dublin, Ireland: Dublin City University and ACL, 2014: 2335-2344.

[6] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces[C]//Proceedings of joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island, Korea: ACL, 2012: 1201-1211.

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

[8] ZHANG S, ZHENG D, HU X, et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia conference on language, information and computation. Shanghai, China: ACL, 2015: 73-78.

[9] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification [C]//Proceedings of the 54th annual meeting of the association for computational linguistics. Berlin, Germany: ACL, 2016: 207-212.

[10] TONGTEP N, THEERAMUNKONG T. A feature-based approach for relation extraction from Thai news documents [C]//Proceedings of the Pacific Asia workshop on intelligence and security informatics. Bangkok, Thailand: Springer, 2009: 149-154.

[11] 王红斌,李金绘,沈　强,等. 基于最大熵的泰语句子级实体从属关系抽取[J]. 南京大学学报:自然科学版,2017,53(4):738-746.

[12] 黄　勇,覃海伦. 老汉汉老军事词典[M]. 北京:军事谊文出版社,2009.

[13] 何阳宇,晏　雷,易绵竹,等. 融合 CRF 与规则的老挝语军事领域命名实体识别方法[J]. 计算机工程,2020,46(8):297-304.

[14] 黄　冰. 老挝语汉语词典[M]. 昆明:国际关系学院昆明分部,2000.

[15] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd international conference on learning representations. CA, USA: ICIR, 2015: 1-15.

[16] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[17] ZHANG D X, WANG D. Relation classification via recurrent neural network[J/OL]. arXiv preprint arXiv: 1508.01006, 2015.