

基于多级注意力融合机制的藏文实体关系抽取

王丽客^{1,2}, 孙媛^{1,2}, 刘思思^{1,2}

(1. 中央民族大学信息工程学院, 北京 100081; 2. 中央民族大学国家语言资源监测与研究少数民族语言中心, 北京 100081)

摘要: 与中英文相比, 藏文实体关系训练语料规模较小, 传统有监督的学习方法难以获得较高的准确率。针对基于远程监督的实体关系抽取存在错误标记的问题, 利用远程监督方法将知识库与文本对齐, 构建藏文实体关系抽取的数据集, 提出一个基于多级注意力融合机制的藏文实体关系抽取模型。在词级别引入自注意力机制来提取单词的内部特征, 在句子级别引入注意力机制为每个实例分配权重, 从而充分利用包含信息的句子, 减少噪声实例的权重。同时引入联合评分函数, 修正远程监督的错误标签, 并将神经网络与支持向量机结合, 实现藏文实体关系分类。实验结果表明, 提出的模型有效提高了藏文实体关系抽取的准确率, 且优于基线模型效果。

关键词: 藏文; 实体关系抽取; 多级注意力融合机制; 支持向量机

中图分类号: TP3

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202146

Tibetan entity relation extraction based on multi-level attention fusion mechanism

WANG Like^{1,2}, SUN Yuan^{1,2}, LIU Sisi^{1,2}

1. School of Information Engineering, Minzu University of China, Beijing 100081, China

2. National Language Resource and Monitoring and Research Center of Minority Languages, Minzu University of China, Beijing 100081, China

Abstract: Compared with Chinese and English, the training corpus of Tibetan entity relation is smaller, so it is difficult to obtain higher accuracy based on traditional supervised learning methods. And there exists the problem of wrong labels in distant supervision for relation extraction. To solve these problems, the distant supervision method was used to construct the data set of Tibetan entity relation extraction through aligning the knowledge base with texts, which could alleviate the problem of lacking of large-scale corpus in Tibetan. And a Tibetan entity relation extraction model based on multi-level attention fusion mechanism was proposed. The self-attention was added to extract internal features of words in word level. The selective attention mechanism could assign weights of each instance, so as to make full use of informative sentences and reduce weights of noisy instances. Meanwhile, a joint score function was introduced to correct wrong labels, and neural network was combined with support vector machine to extract relations. Experimental results show that the proposed model can effectively improve the accuracy of Tibetan entity relation extraction, and is better than the baseline.

Key words: Tibetan, entity relation extraction, multi-level attention fusion mechanism, support vector machine

1 引言

实体关系抽取从无结构化文本中自动抽取实体对之间的语义关系, 并构建结构化的知识, 为知

识图谱、问答系统等研究提供重要的支撑。相比于语料丰富的中英文领域, 藏文实体关系训练语料规模较小, 传统有监督的学习方法难以获得较高的准确率。为了解决训练语料不足的问题, Mintz M 等

收稿日期: 2021-02-01; 修回日期: 2021-02-23

通信作者: 孙媛, tracy.yuan.sun@gmail.com

基金项目: 国家自然科学基金资助项目 (No.61972436)

Foundation Item: The National Nature Science Foundation of China (No.61972436)

人^[1]提出了利用远程监督自动生成训练语料的方法,其主要思想是对齐知识库和文本,将对齐后的句子构建成实体关系抽取的训练数据,并用其训练关系抽取器,预测文本中实体对之间的关系。

基于远程监督的方法可以为藏文实体关系抽取自动标注大规模的数据,然而,在远程监督方法中,假设对于知识库中的三元组 $\langle e_1, r, e_2 \rangle$,文本中所有出现了实体 e_1 和 e_2 的句子都被认为是关系 r 的训练数据,这样会导致假设性太强,不可避免地出现错误标记的问题,影响实体关系抽取的准确率。

本文基于远程监督方法,将构建的藏文知识库与文本对齐,自动标记数据,训练关系抽取模型,实现藏文实体关系抽取。为了减少错误标记的影响,本文提出了基于多级注意力融合机制的藏文实体关系抽取模型,在词级别引入自注意力机制来提取单词的内部特征,在句子级别引入注意力机制为每个实例分配权重,从而充分利用包含信息的句子,减少噪声实例的权重。同时引入联合评分函数,修正远程监督的错误标签,并将神经网络与支持向量机(support vector machine, SVM)结合,实现藏文实体关系分类。

2 实体关系抽取的相关工作

实体关系抽取任务的目标是识别出句子内实体对之间的关系,这是自然语言处理的关键任务之一。随着机器学习的发展,基于深度学习的实体关系抽取成为热点。基于深度学习的方法可以自动提取和学习数据特征,且能较好地利用关系的统计特征。Liu C Y 等人^[2]利用卷积神经网络(convolutional neural network, CNN)进行分类,融合相近词的词向量和词法特征对句子进行编码,但是该模型忽略了词向量的语义信息,没有做到端到端的关系抽取。Zeng D J 等人^[3]使用包含池化层的卷积神经网络模型进行关系分类,融合预训练的词向量和实体的位置特征。Nguyen T H 等人^[4]在卷积层引入多个尺寸的卷积核提取特征,F1 值达到 82.8%。卷积神经网络在关系抽取上效果很好,但是其网络架构决定了其只能学到序列的局部特征,无法学到长距离依赖的信息。

循环神经网络(recurrent neural network, RNN)的提出解决了长距离依赖的问题,其能够同时获取序列之前和之后的信息。Zhang D X 等人^[5]使用双向循环神经网络(bidirectional recurrent neural

network, BiRNN)来建模,在词汇特征上融入实体的位置标签特征。但是双向循环神经网络容易出现梯度爆炸和梯度消失的问题,无法学到更多的长距离依赖信息。为了解决这个问题,Zhou P 等人^[6]使用基于 LSTM 的双向循环神经网络对句子进行建模,使用注意力机制学习句子的语义信息,证明了注意力机制的有效性。Wang L L 等人^[7]、Zhu J Z 等人^[8]利用注意力机制提高了实体关系的准确性。

为了解决语料库不足的问题,Mintz M 等人^[1]提出了远程监督的方法,将文本与预先构建好的知识库对齐,实现对话料库的自动标注。Dietterich T G 等人^[9]提出了基于多示例学习的方法来减少错误标记,采用 NYT+Freebase 数据集,将包含相同实体的句子分成一个包,以包为单位提取出更有效的特征。为了保证包内学到的特征对关系是有价值的,Riedel S 等人^[10]在此基础上假设在包含实体对的包中至少有一个句子表达该关系,并引入无向图模型进行预测。随着深度神经网络的发展,Zeng D J 等人^[11]使用分段卷积神经网络来自动提取句子的特征,但他们假设每个包中只有一个句子是正例,这样会忽略未被选中的句子信息。针对这一问题,Lin Y K 等人^[12]利用注意力机制给句子分配权重。参考文献[13-15]的研究表明,注意力机制在实体关系抽取任务中取得了良好的效果。然而,当前的注意力机制认为词对句子表征的贡献是相等的,忽视了词在句子中对关系类别的作用。

上述研究都是针对中英文语料的,对于藏文,目前没有公开的实体关系训练语料。尽管笔者在前期构建了一定规模的藏文实体关系训练语料并进行了实验^[16-18],但相比中英文大规模语料库,规模太小,数据稀疏,实验结果并不理想。远程监督用于关系抽取的方式是用预先构建的知识库来对齐朴素文本,自动标注数据,在一定程度上减少了人工标注的成本,缓解了藏文训练语料不足的问题。

3 基于多级注意力融合机制的藏文实体关系抽取方法与模型

基于远程监督的方法为藏文实体关系抽取自动标注大规模的数据,但是在此过程中,不可避免地出现错误标记的问题,影响实体关系抽取的准确率。为进一步减少错误标记,本文提出了基于多级注意力融合机制的藏文实体关系抽取模型,动态减少噪声实例的影响,同时在输出层输出软标签

(soft-label), 结合神经网络与 SVM 方法实现藏文实体关系分类, 基于多级注意力融合机制的藏文实体关系抽取模型整体框架如图 1 所示。其中, 双向长短期记忆 (bidirectional long short term memory, BiLSTM) 神经网络在正向 LSTM 的基础上增加了反向 LSTM, 并将两者结合, 这样就能在指定的时间范围内有效地使用过去的特征 (通过前向状态) 和未来的特征 (通过后向状态)。

3.1 远程监督构建藏文实体关系训练语料

本文将构建的藏文知识库与藏文文本语料对齐, 自动标记数据, 训练关系抽取模型, 实现藏文实体关系抽取。本文利用构建的藏文知识库 (三元组 92 950 条, 关系 4 308 条), 与大规模藏文文本语料对齐。若一个句子中出现知识库中的两个实体, 则将句子中的两个实体之间的关系表示为三元组中标记的关系, 从而自动标注大规模的数据, 减少人工标记。但是, 远程监督的假设性太强, 会导致错误标记问题。例如, 知识库中有实体对 <ཆོ་ཤ་སེ (乔布斯), ལཱ་ཁྲཱི (苹果)>, 语料库中包含该实体对的句子可能多个, 如 ཆོ་ཤ་སེ་ནི་ལཱ་ཁྲཱི་ཁྲིའི་ཁྲིའི་འཛིན་པ་ཡིན། (乔布斯是苹果公司的总裁) 和 ཆོ་ཤ་སེ་ལཱ་ཁྲཱི་བློ་བྱ་དགའ། (乔布斯喜欢吃苹果)。远程监督方法会将所有包含实体对 <ཆོ་ཤ་སེ (乔布斯), ལཱ་ཁྲཱི (苹果)> 的句子都提取出来作为训练语句。在上述描述中第 1 句确实表示/people/company/founders 的关系, 而第 2 句中的两个实体只是同时出现, 并没有表示该关系。

因此, 本文使用远程监督模型——分段卷积神经网络 (piecewise convolutional neural network, PCNN) [11], 同时加入句子级别的注意力机制和词级别的注意力机制, 动态减少噪声实例的影响。

3.2 句子级别的注意力机制

为了准确提取特征向量包含的信息, 减少错误标记的噪声问题, 本文在编码层将 ELMo 语言模型^[19]与 PCNN 融合, 根据不同的语境为单词生成不同的词向量, 然后引入多示例学习的方法, 根据实体将输入语句对划分为包, 其目的是预测未出现过的包的标签。在实验过程中, 多示例学习确实起到了一定的修正错误标签的作用, 但是其前提假设是具有相同实体对的句子集中, 至少有一个句子包含知识库中的关系, 并且只为每个包挑出一个最可能的句子来训练, 这会丢失大量包含在被忽略的句子中的丰富信息。所以为了能够充分利用包内的所有信息句, 有效地减少错误标记实例的影响, 本文引入了句子级别的注意力机制来解决这个问题。基于句子级别的注意力机制的 PCNN 架构如图 2 所示, 其中, t_i 表示多示例学习中每个包内的句子, α_i 是每个句子的权重值。

本文使用选择性注意来弱化噪音句子, α_i 的定义如式 (1) 所示:

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)} \quad (1)$$

其中, e_i 、 e_k 分别是预测输入的第 i 个和第 k 个句子关系的置信度得分函数, 计算过程如式 (2) 所示, 其中, $t_i^{q_i}$ 为 q_i 个实体对应的第 i 个句子, A 为加权对角矩阵, r 为关系向量。

$$e_i = t_i^{q_i} \cdot A \cdot r \quad (2)$$

在学习到参数的权重之后, 本文将包的向量表示送入全连接神经网络来预测关系的概率分布, 其中 t_{q_i} 和 r_{q_i} 表示第 q_i 个实体对的原始句子及其对应的句子表示。

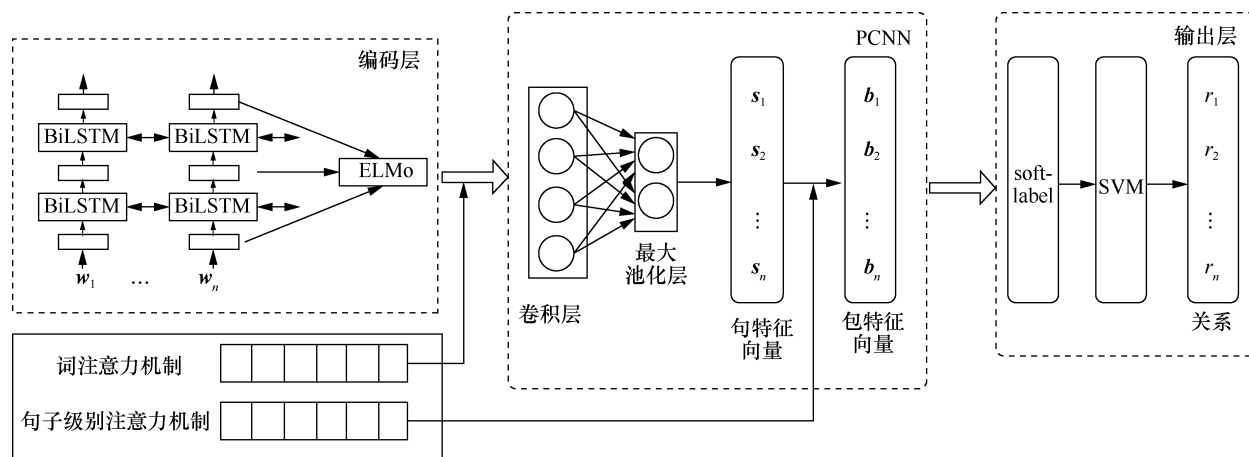


图1 基于多级注意力融合机制的藏文实体关系抽取模型整体框架

3.3 词级别的注意力机制

本文发现句子级别的注意力机制提高了藏文实体关系抽取的准确率。但是,在实验中发现,并不是句子中的所有单词都有助于表达给定的关系,而且句子中的关键词或短语确实可以提高实体关系抽取任务的性能。前人的研究成果证明了深度学习模型中的门控循环单元(gated recurrent unit, GRU)神经网络因其独特的网络结构可以充分学习单词的上下文信息。因而,本文使用双向门控循环单元(bidirectional gated recurrent unit, Bi-GRU)的词注意力机制,同时引入自注意力机制,进一步提高了关系抽取模型的准确率。

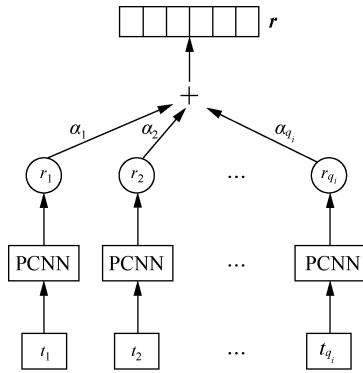


图2 基于句子级别的注意力机制的PCNN架构

Bi-GRU 在一个句子中同时朝前和朝后两个方向运行,以捕获单词上下文的两边。基于 Bi-GRU 的词注意力机制模型的隐藏层包含3个部分:Bi-GRU、词注意力机制和全连接层,模型结构如图3所示。

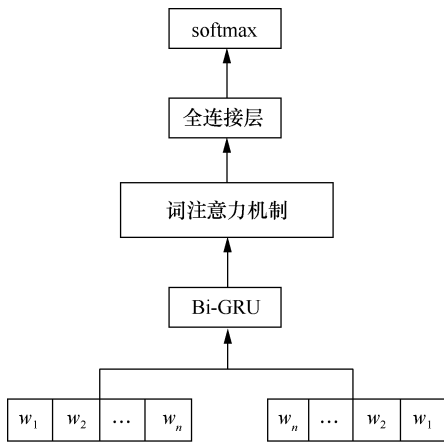


图3 基于 Bi-GRU 的词注意力机制模型结构

在一个句子 s_i 中可能只有几个关键词与所表达的关系有关。在模型中,本文通过词注意力机制,用权重来表示单词与关系的重要程度。句中的每个单词 $x_{i,j}$ 通过 Bi-GRU 神经网络生成正向隐表示 $\vec{h}_{i,j}$

和反向隐表示 $\overleftarrow{h}_{i,j}$ 。每个单词都可以表示为正向和反向隐表示的串联融合,具体计算如式(3)~式(5)所示。其中, $w_{i,j}$ 为 GRU 正向输出结果 $\vec{h}_{i,j}$ 与反向输出结果 $\overleftarrow{h}_{i,j}$ 的串联融合。 $u_{i,j}$ 表示包内第 j 个单词与第 i 个句子的相关性, A 是权重矩阵, r 是关系序列向量,用来计算一个单词与关系向量的相关性。注意力权重值 $\alpha_{i,j}$ 通过 softmax 函数计算得到。在生成句子向量时,不仅将单词的上下文进行加权求和,而且为了更有效地学习句子的深层特征,在输入部分将词注意力机制与上文的 PCNN 融合,在实体对之前、之间和之后分别应用分段最大池化操作,生成分段最大池化句向量,最后通过一个 tanh 非线性层得到每个关系的概率。

$$w_{i,j} = [\vec{h}_{i,j}, \overleftarrow{h}_{i,j}] \quad (3)$$

$$u_{i,j} = w_{i,j} \times A \times r \quad (4)$$

$$\alpha_{i,j} = \frac{\exp(u_{i,j})}{\sum_{l=1}^M \exp(u_{i,l})} \quad (5)$$

本文在构造词注意力机制模型时引入了自注意力机制。在自注意力机制中,通过计算所有单词的重要程度,学习单词的内部依赖关系,捕捉单词的深层特征来表达序列,从而捕捉到序列中术语之间的复杂交互。自注意力机制通过引入缩放点积 attention() 来实现。

其中, Q 、 K 、 V 是来自相同输入的3个矩阵,首先要计算 Q 与 K 之间的点积,以得到 Q 与 K 的相似度。在计算相似度时,为了防止点积的结果过大,会通过除以 $\sqrt{d_k}$ 进行调节,最后使用 softmax 操作将结果归一化为概率分布,然后再乘以矩阵 V 就可以得到权重求和的表示,如式(6)所示:

$$\text{attention}(K, Q, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

计算藏文句子中每一个单词相对于其他单词的权重,可以优化单词的向量表示。并且在计算的过程中,一般单词对本身的注意力权重是最大的。

3.4 联合评分函数

注意力机制依然使用远程监督标记的方法进行降噪,为进一步缓解标签错误的问题,在训练过程中使用联合评分函数来动态修正包的标签。测试

过程仍然使用远程监控方法标记的标签。软标签方法指的是在训练阶段使用相似的语义结构，进而纠正错误的远程监督标签，换句话说就是同样的包在不同的训练阶段可以被标记为不同的标签。将基于实体对表示的关系分数与远程监督标签的置信度结合，得到一个新的标签，即针对某一实体对的标签，根据正确的实例标签动态更正每个包的标签，如式（7）所示：

$$r_i = \arg \max(\mathbf{o} + \max(\mathbf{o})\mathbf{A} \odot D_i) \quad (7)$$

其中， D_i 表示远程监督方法生成的标签， $\mathbf{A}$ 表示远程监督标签的置信度向量， $\odot$ 表示点积运算， $\mathbf{o}$ 表示基于实体对的关系得分向量， $\max(\mathbf{o})$ 是缩放常数。第 i 个关系类别的得分如式（8）所示：

$$o_i = \frac{\exp(N \cdot e_i + b)}{\sum_k \exp(N \cdot e_k + b)} \quad (8)$$

其中， e_i 和 e_k 分别是预测输入的第 i 个和第 k 个句子关系的置信度得分函数， N 为关系矩阵， b 是偏置向量。在训练过程中，将交叉熵作为损失函数，如式（9）所示，其中， p 表示求概率， $J(\theta)$ 为损失函数值， θ 为神经网络参数。

$$J(\theta) = \sum_{i=1}^n \log p(r_i | e_i; \theta) \quad (9)$$

最后，本文将 SVM 分类器与神经网络最后的输出结合，并在输出层转化为一个多分类的结构，以提高实体关系抽取的准确率。

4 实验对比分析

4.1 数据集

本文预先构建了藏文知识库，将其与藏文文本进行对齐，获得语料库。本文主要将职业、民族、性别、父亲、母亲、出生地、死亡日期、出生日期、NA（表示实体对没有关系）9种关系的语料作为数据集，共 22 156 个句子。其中 17 522 个句子被用于训练实体关系抽取模型，4 634 个句子被用于测试，将语料库中数据的格式统一为六元组，具体的格式见表 1。

4.2 评测指标

为有效验证语料库的效果以及模型的性能，实验将 TOP- N 准确率、 P - R 值和 F1 值作为评价指标。

（1）TOP- N 准确率

测试语料以句子为单位进行关系预测，由于测

试语料的句子数量是不确定的，为验证模型的稳定性，从测试集的新实体对中选择预测的关系概率值最高的前 N 个，然后人工检查其中标记正确的实例，计数为 n ，则查准率为 n/N 。为验证实验结果的有效性，本文将前 100 个、前 200 个和前 300 个实例的准确率进行比较， $P@avg$ 是前 100 个、前 200 个和前 300 个实例的准确率之和的平均值。

（2） P - R 值和 F1 值

召回率 R 表示所有的正例中，预测正确实例所占的比例。准确率 P 表示预测的结果中，正确的结果所占的比例，本文定义如下：如果正确地将正实例分类为正类，表示为 TP；如果将正实例分类为负类，表示为 FN；如果正确地将负实例分类为负类，表示为 TN；如果将负实例分类为正类，表示为 FP。计算式如式（10）~（12）所示：

$$P = \frac{TP}{TP + FP} \quad (10)$$

$$R = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = \frac{2PR}{P + R} \times 100\% \quad (12)$$

4.3 实验设置与结果分析

模型的参数设置见表 1 和表 2。

表 1 基于 Bi-GRU 的词注意力机制模型部分参数设置

参数名	参数值
Batchsize	50
Learning_rate	0.1
Dropout	0.5

表 2 自注意力机制模型部分参数设置

参数名	参数值
Num_blocks	4
Num_heads	10
Hidden_size	230

本文用 F1 值和前 100 个、前 200 个、前 300 个实例的准确率以及 $P@avg$ 来评估模型，实验结果见表 3、表 4 和表 5。

（1）句子级别的注意力机制实验结果

本文对比了 CNN 和 PCNN 的神经网络结构，结果见表 3。其中，AVE 是平均注意力机制，ATT 是选择性注意力机制，以 CNN+AVE 为基线模型。

表3 不同模型句子级别的注意力机制的结果对比

模型	F1	P@100	P@200	P@300	P@avg
CNN+AVE	32.6%	47.5%	36.2%	32.5%	38.8%
CNN+ATT	36.2%	52.4%	39.8%	33.8%	42.0%
PCNN+AVE	37.9%	45.0%	42.5%	39.2%	42.2%
PCNN+ATT	42.7%	48.6%	44.0%	41.7%	44.8%

从表3可以看出：①对于CNN和PCNN架构来说，ATT方法在测试时都达到了很好的性能，这说明了句子级别的注意力机制在多示例学习中的有效性；②在p@avg方面，CNN+ATT相比CNN+AVE有3.2%的改进，PCNN+ATT相比PCNN+AVE有2.6%的改进，这表明，利用包内所有的句子会带来更多的信息，当然也可能会带来一些额外的噪音；③PCNN的效果比CNN好，说明分段卷积确实可以学到每个部分更多的信息，有利于关系抽取任务。

(2) 词级别的注意力机制实验结果

在实验过程中，本文对词注意力机制进行两种方式的评价：定量分析和定性分析。定量分析是指对注意力机制的贡献进行量化，即结合下游的实体关系抽取任务。定性分析是将注意力机制进行可视化，为整个句子构建一个热力图，热力图颜色深浅与每个单词的注意力得分成正比。也就是说，词的贡献越大，颜色越深。

定量分析如下。

将两种词注意力机制与下游的实体关系抽取任务相结合，分析其效果。从表4可以看出：加入词注意力机制提高了藏文实体关系抽取任务的准确率，说明单词对于关系的分类确实起到一定的作用。其中，加入自注意力机制效果后，相比PCNN+ATT模型，F1值提高了7.8%，P@avg提高了8.4%，这说明自注意力机制在一个句子中可以同时朝前和朝后两个方向运行，以捕获一个单词在前文和后文的特征，较好地将上下文语境融入其中。

定性分析如下。

将两种词注意力机制进行可视化，更加直观地判

断词注意力机制的效果（本文使用的例子均为虚拟）。

由图4可知，在进行关系分类任务时，句子中的每个词对于关系类别的重要性都不同。从图4可以看出，“卓玛”是人名，在注意力机制中对父亲和母亲等关系分配了较大的权重，“1968”这样的时间名词则对出生日期和死亡日期等关系有更多的贡献，地点名词很明显能够表达出生地的关系，其中“出生于”一词可以帮助这个句子识别出生地这一关系，属于关键词类别，分配到了0.73的较大权重。

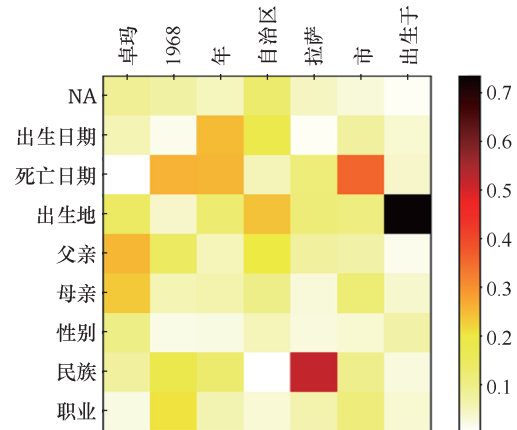


图4 传统词注意力机制热力图

从图5可以看出，自注意力机制可以计算出一个句子内单词与单词之间的注意力得分，并且每个单词对自身的权重是最大的。其中，“卓玛”对“拉萨”和“出生于”等词的影响较大，这也符合人们的行为，在神经网络学习特征的过程中，要符合现实生活中人类的行为，大规模的语料库可以使机器更好地学到与人类行为有关的词。“拉萨”对“市”的影响较大，因为在句中两个词一般同时出现。“出

表4 基于Bi-GRU的注意力机制模型与自注意力机制模型结果对比

模型	F1	P@100	P@200	P@300	P@avg
PCNN+ATT	42.7%	48.6%	44%	41.7%	44.8%
+Bi-GRU+ATT	47.1%	54.4%	48.2%	46.8%	49.8%
+Self-ATT	50.5%	56.9%	52.6%	50.1%	53.2%

生于”一词则对一些地名标志影响较大,这说明该动词后一般会出现地名。从句子整体来看,自注意力机制可以计算出单词与单词之间的权重得分,有效地学到句内信息,提取单词的内部特征。

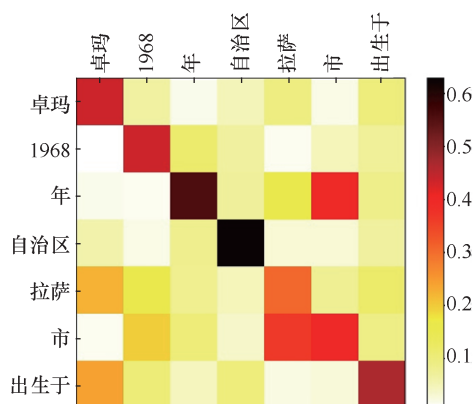


图5 自注意力机制热力图

通过图4和图5两个热力图可以很直观地看到,句子中每个单词对于关系的重要程度,可以很好地识别出句中的关键词,也能看到句子内部单词与单词之间的影响,实验结果也证明了本文提出的模型的有效性。

在表4中+Self-ATT模型的基础上,添加了软标签方法和SVM分类器,实验结果见表5,可以得到如下结论。

(1) 在藏文语料中,加入软标签方法,实验结果比表4中+Self-ATT模型的F1值提高了5.4%,P@avg值提高了1.1%,这证明软标签方法具有一定的去噪效果。

(2) SVM分类器与神经网络相结合确实取得了更好的效果,在藏文上F1值为62.6%,说明利用远程监督方法确实可以缓解人工标注数据集,而且模型效果也验证了其改进是有效的,具有更高的精度。

表5 添加软标签方法和SVM分类器的实验结果

模型	F1	P@100	P@200	P@300	P@avg
+soft-label	55.9%	59.8%	53.2%	49.9%	54.3%
+SVM	62.6%	64.1%	59.7%	53.8%	59.2%

5 结束语

本文主要通过构建基于多级注意力融合机制的藏文实体关系抽取模型来减少错误标记,并在远程监督方法自动生成的语料集上进行了对比实验。

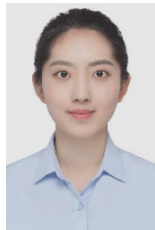
首先,模型利用融合多示例学习的PCNN方法,然后在句子级别的注意力机制的基础上,融合词级别的注意力机制,并分别在句子级别和单词级别的注意力机制中进行对比实验。实验结果证明,词注意力机制可以学到句中单词的内部信息,捕捉句子的内部结构。此外,本文还在此模型基础上,利用SVM分类器优化传统神经网络的softmax输出。实验结果证明,本文提出的方法确实可以提高藏文实体关系抽取任务的准确率,而且模型在拟合新数据方面的能力有了较大的提高。未来还需要研究如何使语料的质量更高,更加符合人工处理的标准,并且从实际场景出发,构建更加有效而且鲁棒性更好的实体关系抽取模型。

参考文献:

- [1] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLPJ. [S.l.:s.n.], 2009: 1003-1011.
- [2] LIU C Y, SUN W B, CHAO W H, et al. Convolution neural network for relation extraction[C]//Proceedings of the 9th International Conference on Advanced Data Mining and Applications, Part II. [S.l.:s.n.], 2013: 231-242.
- [3] ZENG D J, LIU K, LAI S W, et al. Relation classification via convolutional deep neural network[C]//Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. [S.l.:s.n.], 2014: 2335-2344.
- [4] NGUYEN T H, GRISHMAN R. Relation extraction: perspective from convolutional neural networks[C]//Proceedings of NAACL-HLT. [S.l.:s.n.], 2015: 39-48.
- [5] ZHANG D X, WANG D. Relation classification via recurrent neural network[J]. arXiv preprint, 2015, arXiv:1508.01006.
- [6] 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. [S.l.:s.n.], 2016: 207-212.
- [7] WANG L L, CAO Z, MELO G D, et al. Relation classification via multi-level attention CNNs[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. [S.l.:s.n.], 2016: 1298-1307.
- [8] ZHU J Z, QIAO J Z, DAI X X, et al. Relation classification via target-concentrated attention CNNs[C]//Proceedings of the 31st International Conference on Neural Information Processing. [S.l.:s.n.], 2017: 137-146.
- [9] DIETTERICH T G, LATHROP R H, LOZANO-PEREZ T. Solving the multiple instance problem with axis-parallel rectangles[J]. Journal of the Artificial Intelligence, 1997, 89(1-2):31-71.
- [10] RIEDEL S, YAO L M, MCCALLUM A. Modeling relations and their mentions without labeled text[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases.

- [S.l.:s.n.], 2010: 148-163.
- [11] ZENG D J, LIU K, CHEN Y B, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. [S.l.:s.n.], 2015: 1753-1762.
- [12] LIN Y K, SHEN S Q, LIU Z Y, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. [S.l.:s.n.], 2016: 2124-2133.
- [13] JIANG X T, WANG Q, LI P, et al. Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks[C]//Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. [S.l.:s.n.], 2016: 1471-1480.
- [14] FENG X C, GUO J, QIN B, et al. Effective deep memory networks for distant supervised relation extraction[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. [S.l.:s.n.], 2017: 4002-4008.
- [15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. [S.l.:s.n.], 2017: 5998-6008.
- [16] 朱臻, 孙媛. 基于 SVM 和泛化模板协作的藏语人物属性抽取[J]. 中文信息学报, 2015, 29(6):220-227.
- ZHU Z, SUN Y. Tibetan person attribute extraction based on SVM and pattern[J]. Journal of Chinese Information Processing, 2015, 29(6): 220-227.
- [17] 夏天赐, 孙媛. 基于联合模型的藏文实体关系抽取方法研究[J]. 中文信息学报, 2018, 32(12): 76-83.
- XIA T C, SUN Y. Tibetan entity relation extraction based on joint model[J]. Journal of Chinese Information Processing, 2018, 32(12): 76-83.
- [18] 郭莉莉, 孙媛. 基于 BP 神经网络的藏语实体关系抽取[J]. 软件导刊, 2019, 18(3): 7-9, 15.
- GUO L L, SUN Y. Tibetan entity relation extraction based on BP neural network[J]. Software Guide, 2019, 18(3): 7-9, 15
- [19] MATTHEW E P, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.:s.n.], 2018: 2227-2237.

[作者简介]



王丽客（1996—），女，中央民族大学信息工程学院硕士生，主要研究方向为自然语言处理、知识图谱。



孙媛（1979—），女，博士，中央民族大学信息工程学院副教授，主要研究方向为自然语言处理、知识图谱、问答系统。



刘思思（1998—），女，中央民族大学信息工程学院硕士生，主要研究方向为自然语言处理、知识图谱。