◆人工智能算法与应用◆

基于字词混合和 GRU 的科技文本知识抽取方法*

欧阳苏宇,邵蓥侠**,杜军平,李　昂
(北京邮电大学计算机学院,智能通信软件与多媒体北京重点实验室,北京　100082)

摘要:知识抽取任务是从非结构化的文本数据抽取三元组关系(头实体-关系-尾实体)。现有知识抽取方法分为流水式方法和联合抽取方法。流水式方法将命名实体识别和实体知识抽取分别用各自的模块抽取,这种方式虽然有较好的灵活性,但训练速度较慢。联合抽取的学习模型是一种通过神经网络实现的端到端的模型,同时实现实体识别和知识抽取,能够很好地保留实体和关系之间的关联,将实体和关系的联合抽取转化为一个序列标注问题。基于此,本文提出了一种基于字词混合和门控制单元(Gated Recurrent Unit,GRU)的科技文本知识抽取(MBGAB)方法,结合注意力机制提取中文科技资源文本的关系;采用字词混合的向量映射方式,既在最大程度上避免边界切分出错,又有效融入语义信息;采用端到端的联合抽取模型,利用双向 GRU 网络,结合自注意力机制来有效捕获句子中的长距离语义信息,并且通过引入偏置权重来提高模型抽取效果。

关键词:知识抽取　向量映射　GRU　三元组关系　联合抽取方法

中图分类号:TP39　　**文献标识码**:A　　**文章编号**:1005-9164(2022)04-0634-08
DOI:10.13656/j.cnki.gxkx.20220919.003

无论是专业科技资源平台,还是社交媒体场景,都有大量的科技文本数据[1-3],对这些数据进行知识抽取能更好地进行信息挖掘和利用[4,5]。知识抽取任务[6]是从非结构化的文本数据抽取三元组关系(头实体-关系-尾实体),现有知识抽取研究主要基于循环神经网络(Recurrent Neural Network,RNN)。RNN 的同一层节点之间是相互连接的,对于每一个时间步长,都有来自前面时间步长的信息,并加以权重用以控制[7]。长短期记忆网络(Long Short-Term Memory,LSTM)用于解决 RNN 在训练过程中容易出现梯度爆炸和梯度消失的问题[8]。门控制单元(Gated Recurrent Unit,GRU)建立在 LSTM 的基础上,仅由重置门(Reset Gate)和更新门(Update Gate)组成[9]。

现有基于深度学习的知识抽取方法分为流水式方法和联合抽取方法。本文采用联合知识抽取模型,很好地保留实体和关系之间的关联,将实体和关系的联合抽取转化为序列标注问题。为了在最大程度上

收稿日期:2022-04-16
*国家重点研发计划项目(2018YFB1402600)和国家自然科学基金项目(61772083,61877006,61802028,62002027)资助。
【作者简介】
欧阳苏宇(1997-),男,在读硕士研究生,主要从事自然语言处理、数据挖掘和深度学习研究。
【**通信作者】
邵蓥侠(1988-),男,副教授,主要从事大规模图分析、并行计算框架和知识图谱分析研究,E-mail:shaoyx@bupt.edu.cn。
【引用本文】
欧阳苏宇,邵蓥侠,杜军平,等. 基于字词混合和 GRU 的科技文本知识抽取方法[J]. 广西科学,2022,29(4):634-641.
OUYANG S Y,SHAO Y X,DU J P,et al. Knowledge Extraction Method of Scientific and Technological Text Based on Word Mixing and GRU [J]. Guangxi Sciences,2022,29(4):634-641.

避免边界切分出错,选择字标注的方式,即以字为基本单位进行输入。但是在中文中,单纯的字 Embedding 难以存储有效的语义信息,因此为了更有效地融入语义信息,本文设计了一种字词混合方式。同时结合自注意力机制来捕获句子中的长距离语义信息,并且通过引入偏置权重来提高模型抽取效果。

1 相关工作

知识抽取是从文本中抽取结构化信息[10],文本实体关系抽取是信息抽取的一个子域,是指从文本关系提取语义关系,这种语义关系存在于实体对之间。定义文本 S,关系集合 $R=\{r_1,r_2,\cdots\}$,文本关系抽取就是根据 S 和 R 抽取三元组 (h,r,t) 的过程,其中,h 表示头实体,t 表示尾实体,r 表示 h 与 t 之间的关系[11]。一般来说,文本关系抽取的步骤分为命名实体识别(Named Entity Recognition,NER)[12]和关系分类(Relation Extraction,RE)[13]。在深度学习中,文本关系抽取主要基于有监督和远程监督两种方法进行研究。有监督的关系抽取方法主要包括流水式学习和联合学习两种。

流水式抽取通常将命名实体识别和语义关系分类独立进行,先识别出文本数据中存在的实体,再根据实体对判断之间的关系是否存在。早期的流水式学习方法主要基于卷积神经网络(Convolutional Neural Networks, CNN)[14,15]。Zeng 等[16]利用 CNN 网络进行关系分类,设计基于词法级别(lexical level)和句子级别(sentence level)的特征提取网络,对两种特征进行融合得到编码向量,在编码器网络后接全连接层,利用 softmax 进行关系分类,在公开数据集上验证了方法的有效性。在此基础上,Nguyen 等[17]加入了多尺寸卷积核,完全使用句子级别特征,能自动学习句子中的隐含特征。近年来,许多学者基于 RNN 开展了研究。Socher 等[18]首次采用 RNN 对文本进行句法解析,句法解析树上的每个节点由向量和变换矩阵两部分组成,对于任意句法类型和长度的词语和句子,均可以学习其组合向量表示。LSTM 是一种特殊的 RNN。Xu 等[19]提出了采用 SDP-LSTM 模型进行关系分类,基于最短依赖路径的思想,过滤文本无用信息,使用 LSTM 将异构信息进行有效集成,制定有效的 dropout 策略防止出现过拟合的可能。Zhang 等[20]提出了双向递归卷积神经网络模型(BRCNN),基于最短依赖路径,使用了双向长短期记忆网络(Bi-LSTM),考虑实体之间关系的方向性,利用词语前后的信息进行关系抽取。

流水式方法将命名实体识别和关系分类视为两个独立的任务,忽略了任务之间存在的关系。同时,由于任务之间存在先后顺序,导致实体识别的误差影响到后续关系分类,出现误差传播现象。另外,由于缺乏关系的实体对无法两两配对,导致在关系分类中带来多余信息,出现实体对冗余的现象。近年来,许多研究尝试将两个任务联合进行学习,即将命名实体识别和关系分类融合为单个任务。联合学习方法主要包括基于参数共享的方法、基于序列标注的方法和基于图结构的方法。

①基于参数共享的方法是在命名实体识别和关系分类两个任务中间设计共享编码层,在训练中得到最佳的全局参数。Miwa 等[21]首次设计基于 LSTM 的共享编码层,以获得单词序列和句法依存树上的子结构信息。Zheng 等[22]在此基础上,设计基于向量嵌入层和 Bi-LSTM 层的共享编码层,命名实体名识别模块采用 LSTM,关系分类模块采用 CNN,有效捕获了长文本实体标签之间的距离依赖关系。

②基于序列标注的方法是将命名实体识别和关系分类两个任务融合成序列标注的问题。为了解决基于参数共享的方法容易产生实体对冗余的问题,Zheng 等[23]提出了一种新的标注方案,将联合提取任务转换为标注问题,研究了不同端到端的模型[24]的关系抽取性能,另外使用偏置损失函数来增强相关实体之间的关联。Bekoulis 等[25]利用条件随机场(Conditional Random Field,CRF)将命名实体识别和关系抽取任务建模为一个多头选择问题(Multi-Head Selection Problem),将关系分类任务看作多个二分类任务,从而使得每个实体能够与其他所有实体判断关系,有效解决了关系重叠的问题。

③基于图结构的方法是利用图神经网络(Graph Neural Networks,GNN)对关系进行抽取。Wang 等[26]提出基于图结构的联合学习模型,同时使用偏置权重的损失函数削弱无效标签的影响,增强了相关实体间的关联。此外,Fu 等[27]提出了一种端到端的关系抽取模型 GraphRel,通过关系加权图卷积网络(Graph Convolutional Network,GCN)来考虑实体和关系之间的交互,将 RNN 和 GNN 结合起来提取每个单词的顺序特征和位置依赖特征,有效解决实体对重叠的问题。

基于远程监督的方法可以极大降低人工成本,耗时短,而且领域可移植性强。Mintz 等[28]提出了 DS-

logistic 模型,利用外部知识库,将大规模知识图谱与文本关联,对远程监督标注的数据提取文本语义特征,将其表征输入到分类器中进行关系分类。但是,由于采用自动标注,训练集数据出错的可能性大幅增高,导致目前远程监督实体关系抽取准确率偏低。

2 基于字词混合及 GRU 的科技文本知识抽取(MBGAB)方法

为提高抽取性能,同时解决基于流水式的方法所带来的冗余信息的问题,本节将三元组的抽取问题转变为多序列标签分类问题,设计了编码器-解码器的神经网络结构,使用端对端的模型预测输入文本序列的三元组标签序列。具体来讲,本文在向量序列层采用字词混合向量映射的方式,改善中文分词边界出错可能带来歧义的问题。采用门控循环单元机制对输入句子进行编码,引入自注意力机制来捕获句子中的长距离语义信息,使用带有偏置权重的目标函数来增强实体标签的相关性和降低无用标签的影响度。基于字词混合的文本关系联合抽取方法整体架构如图 1 所示。

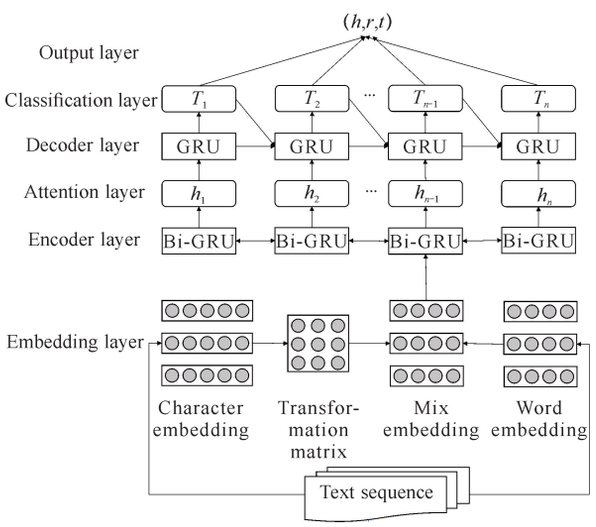

图 1 MBGAB 方法架构

Fig. 1 MBGAB method architecture

2.1 向量序列层

在向量序列层中,针对中文文本单纯的字向量难以存储有效的语义信息,同时为了最大程度上避免边界切分出错,本文设计了一种字词结合的向量混合映射方式。具体来讲,输入以字为单位的文本序列 S,经过随机初始化的字 Embedding 层,得到字向量序列 $S=(c_1,c_2,\cdots,c_n)$。将文本序列分词,通过 Word2Vec 词向量模型提取对应的词向量,为了将词向量序列与字向量序列对齐,使单个词向量 w_i 重复 k 次,k 即为组成该词的字数,得到词向量序列 $S=(w_1,w_2,\cdots,w_n)$。例如对于“硕士研究生”,将该词所对应的词向量重复 5 次得到对齐的词向量序列。得到对齐的词向量后,将词向量经过一个随机初始化的变换矩阵 T,得到与字向量相同维度的向量,并将两者相加。字词混合向量映射公式为

$$x_i=c_i+w_i, \tag{1}$$

其中,x_i 代表融合后的字向量,字词混合向量即为两者加和。此时,文本序列转换成融合后的字向量序列 $S=(x_1,x_2,\cdots,x_n)$。

2.2 编码层

在得到科技资源文本字词混合向量后,此时一个句子的字向量序列可以表示为 $S=[w_1,w_2,\cdots,w_n]$,其中 w_i 表示该句子中的第 i 个汉字,n 表示该句子由 n 个汉字组成。双向 GRU(Bi-GRU)编码层利用先前的隐藏状态 h_{t-1} 和输入单词序列的字向量表示 w_t,计算每个时间步长更新后的隐藏状态 h_t,具体计算公式如下:

$$z_t=\gamma(W_z w_t+U_z h_{t-1}+b_z), \tag{2}$$

$$r_t=\gamma(W_r w_t+U_r h_{t-1}+b_r), \tag{3}$$

$$\tilde{h_t}=\tanh(W w_t+U h_{t-1} r_t+b), \tag{4}$$

$$h_t=(1-z_t h_{t-1}+z_t \tilde{h_t}), \tag{5}$$

其中,γ 表示激活函数,W 和 U 表示权值矩阵,b 表示偏置量。对于单个汉字 w_t,前向 GRU 层计算 w_t 的上文信息并将其编码为 h_t^f,后向 GRU 层计算 w_t 的下文信息并将其编码为 h_t^b,融合得到最终编码信息 $h_t=[h_t^f+h_t^b]$。经过 Bi-GRU 编码层的处理,最终将词嵌入向量序列 $W=w_1,w_2,\cdots,w_N$ 转化为带有句子语义信息的词向量 $H=\{h_1,h_2,\cdots,h_N\}$。

2.3 自注意力层

本文采用自注意力机制来计算单个汉字与其他汉字之间的关联,减少外部信息的依赖,针对句子中的长距离语义关系进行捕获。自注意力编码层的输入来自 Bi-GRU 编码层的输出,输入为 $H=h_1,h_2,\cdots,h_N$,输出为 $H^*=h_1^*,h_2^*,\cdots,h_N^*$。首先将输入向量经过线性转换获得 3 个向量序列 Q,K,V,注意力计算公式为

$$h_i^*=\sum_{j=1}^{N}a_{ij}v_j=\sum_{j=1}^{N}\mathrm{softmax}(s(q_i,k_j))v_j。 \tag{6}$$

为了保持梯度稳定,对函数除以 $\sqrt{d_K}$,之后利

用softmax激活函数对结果进行归一化分布，即采用缩放点积的类型函数为注意力打分，生成序列H^*为

$$H^* = \text{softmax}(\frac{Q^{\mathrm{T}}K}{\sqrt{d_K}})V。 \tag{7}$$

2.4 解码层

得到综合上下文编码信息的文本序列后，本文采用GRU结构对文本序列解码产生标签序列。对单词w_t进行标注时，解码层的输入为从编码层得到的词向量表示h_t^*，前一个的预测标签表示T_{t-1}以及解码层中的前一个隐藏状态h_{t-1}^d，经过计算输出得到单词w_t的预测标签状态T_t，具体公式如下：

$$r_t^d = \gamma(W_r^d h_t^* + U_r^d h_{t-1}^d + V_r^d T_{t-1} + b_r^d), \tag{8}$$

$$z_t^d = \gamma(W_z^d h_t^* + U_z^d h_{t-1}^d + V_z^d T_{t-1} + b_z^d), \tag{9}$$

$$\widetilde{h_t^d} = \tanh(W^d r_t^d h_t^* + U^d h_{t-1}^d + V^d T_{t-1} + b^d), \tag{10}$$

$$h_t^d = (1 - z_t^d) h_{t-1}^d + z_t^d \widetilde{h_t^d}, \tag{11}$$

$$T_t = \tanh(W_T h_t^d + b_T^d)。 \tag{12}$$

2.5 分类层

本文采用softmax分类器进行标签分类，根据标签预测向量T_t归一化计算实体标签概率。定义单词T_t在所有标签类型上的评分：

$$Y_t = W_Y T_t + b_Y, \tag{13}$$

其中，W_Y是参数矩阵，b_Y是偏置项。通过softmax层计算单词T_t为标签i的概率为

$$p_t^i = \frac{\exp(Y_t^i)}{\sum_{j=1}^{Nt}\exp(Y_t^j)}, i \in 1,\cdots,k, \tag{14}$$

其中，Y_t^j是x_j句子中第t个词的真实标注，N_t表示标签总数。

通过最大化对数似然函数

$$L = \max\sum_{j=1}^{|D|}\sum_{t=1}^{L_j}\begin{pmatrix}\log(p_t^j \mid x_j, \odot) * I(O) \\ +\alpha * \log(p_t^j \mid x_j, \odot) * (1 - I(O))\end{pmatrix}, \tag{15}$$

其中，$|D|$表示科技资源文本训练数据集的大小，L_i是输入句子中单词的长度，p_t^j是归一化后的标签概率分布。$I(O)$用来区分无用的标签“O”和能指示抽取结果的相关标签，即当tag=‘O’时，$I(O)=1$；当tag≠‘O’时，$I(O)=0$。此外，α是偏置权重，用来控制非“O”标签影响度的超参数，α越大，表示模型中有关系的标签的影响度越大。

MBGAB方法的整体步骤如下：

①固定词向量w_{word}不变，对随机初始化的字向量$w_{\text{character}}$进行训练；

②得到字词混合映射向量$w_i = w_{\text{character}} + w_{\text{word}}$；

③通过双向GRU编码层后，将词嵌入向量序列$W = w_1, w_2, \cdots, w_N$转化为带有句子语义信息的词向量$H = \{h_1, h_2, \cdots, h_N\}$；

④自注意力编码层的输入来自Bi-GRU编码层的输出，输入为$H = h_1, h_2, \cdots, h_N$，输出为$H^* = h_1^*, h_2^*, \cdots, h_N^*$；

⑤GRU解码层经过计算输出得到单词w_t的预测标签状态T_t。

3 验证实验

3.1 数据集

为验证模型对专家学者科技资源信息的知识抽取的可行性，使用部分LIC2019中文抽取语料，其中包含19种关系，将数据集分为24 851条训练集和6 212条测试集，其中训练集与测试集中标签占比基本一致，以保证数据的一致性。

3.2 评估指标

为评估所提算法的效果，本文使用准确率(precision)、召回率(recall)以及F1-score指标对知识抽取效果进行评价。

3.3 实验设置

ME-BiGRU：去除注意力机制和权重偏置。

ME-BiGRU-SA：去除权重偏置。

BIGRU-SA-Bias：去除字词混合Embedding。

ME-GRU-CRF：解码层用条件随机场CRF进行解码。

ME-BiGRU-Bias：去除注意力机制。

对于实验的参数根据算法的结构图进行以下设置，在编码层的输入是采用预训练好的Word2Vec模型生成的词向量，词向量维度为300，字向量使用随机初始化的字Embedding层。在模型训练中，固定Word2Vec词向量不变，只优化变换矩阵和字向量。Bi-GRU编码层的维度为300，GRU解码层的维度为600，偏置权重参数α设置为3。

3.4 MBGAB方法的有效性

使用以下算法进行对比。

①FCM：分别进行实体和知识抽取，是一种流水式抽取模型。

②Attention-BILSTM：使用Attention和双向LSTM对关系进行抽取，是一种流水式抽取方法。

③ MultiR:远程监督算法,是一种联合抽取方法。

④Cotype:将实体、关系、文本特征和类型标签嵌入到两个向量空间,是一种联合抽取方法。

实验结果如表 1 所示。

表 1 知识抽取实验结果对比

Table 1 Comparison of knowledge extraction experiment results

算法 Algorithm	准确率(%) Precision (%)	召回率(%) Recall (%)	F1-score (%)
FCM	43.7	25.8	32.4
Attention-BILSTM	45.8	34.6	39.4
MultiR	51.9	43.4	47.3
CoType	55.2	42.6	48.1
MBGAB	65.3	49.1	56.1

从表 1 可以看出,MBGAB 方法在相同数据集上的表现较其他方法有明显提升,证明了该方法的有效性。FCM 方法和 Attention-BILSTM 方法召回率较低,表明传统的流水式方法在面对关系重叠问题时性能不佳。MultiR 和 CoType 的性能较 FCM 和 Attention-BILSTM 方法更好,可能是基于流水式的方法将两个子任务独立执行,而联合抽取方法利用了两者之间的关联性从而尽可能减少误差传播。对比其他传统联合抽取方法,基于本文端到端的模型在准确率和召回率上有了显著提升,考虑到字词混合映射方式和双向 GRU 编码方式文本语义表示的有效性,同时结合自注意力层对数据的良好适应性,最终使得本文方法在 F1 值上总体提升。

3.5 消融学习

本文提出的算法架构中,核心部分是基于字词混合向量嵌入方式、Bi-GRU 编码层、自注意力层以及 GRU 解码层,并且在目标函数上加入了偏置项。为了观察算法核心组成部分对于关系抽取的改善效果,对这些部分进行消融学习。本节使用以下变种作为对比。

①BiGRU-GRU-SA-Bias$_{(c)}$:去除字词混合嵌入方式,采用字嵌入方式,直接将以字为单位的文本序列经过随机初始化的 Embedding 层进行向量嵌入;

②BiGRU-GRU-SA-Bias$_{(w)}$:去除字词混合嵌入方式,采用词嵌入方式,将文本分词后,采用预训练好的 Word2Vec 模型进行向量嵌入;

③ME-BiGRU-GRU:去除注意力机制和偏置权重,以观察两者对算法性能的影响;

④ME-BiGRU-CRF-SA-Bias:解码层采用条件随机场 CRF,将 CRF 应用于预测实体标签序列;

⑤ME-BiGRU-GRU-SA:去除偏置权重,以验证目标函数中加入偏置权重对性能的影响;

⑥ME-BiGRU-GRU-Bias:去除注意力机制,以验证自注意力层对算法性能的影响。

实验结果对比如表 2 所示。从表 2 可以看出,本文方法(MBGAB)与其他 6 个变种相比,召回率和 F1 值均为最高,证明了方法的各组成部分对于性能提升都有一定贡献。

表 2 不同变种与原算法在知识抽取任务上的实验结果对比

Table 2 Comparison of experimental results between different variants and the original algorithm on knowledge extraction tasks

变种序号 No. of variants	算法 Alogrithm	准确率(%) Precision (%)	召回率(%) Recall (%)	F1-score (%)
1	BiGRU-GRU-SA-Bias$_{(c)}$	64.1	47.8	55.0
2	BiGRU-GRU-SA-Bias$_{(w)}$	64.3	47.3	54.5
3	ME-BiGRU-GRU	63.6	43.8	51.9
4	ME-BiGRU-CRF-SA-Bias	64.6	43.9	52.3
5	ME-BiGRU-GRU-SA	67.3	46.9	55.3
6	ME-BiGRU-GRU-Bias	64.1	46.5	53.9
7	MBGAB	65.3	49.1	56.1

对于变种 1 和变种 2,去除了字词混合嵌入方式,采用字嵌入和词嵌入,本文方法在准确率和召回率上都有提高,在 F1 值上分别提高了 1.1% 和 1.6%。考虑到本文方法使用更为适合中文文本的字向量,微调了预训练词向量,使得向量嵌入层学习了预训练模型所带来的丰富的语义信息,同时也减小了中文分词可能带来的错误,保留了字向量的灵活性。因此,字词混合嵌入方式可以有效改善中文文本的语义表示能力。对于变种 3,去除了自注意力机制和偏置权重后,准确率、召回率以及 F1 值均为最低,表明了本文所用到的数据集对于长距离文本语义和关系标签较为敏感,证明了两者对于方法的性能具有一定影响。对于变种 4,当解码层用 CRF 替换后,准确率下降程度较小,但召回率下降程度较大,考虑到 CRF 擅长计算标注的联合概率,而文本语句中相关联的两个实体标签可能距离过长,GRU 能够更好地学习句子中长距离的依赖关系,因此模型性能效果较好。对

于变种 5，去除偏置权重后，准确率有一定提升，但召回率和 F1 值均降低。考虑到加入偏置权重的目标函数对于关系标签较为敏感，使无效标签的影响程度降低，因此提升了方法对于相关联实体的识别能力。对于变种 6，去除自注意力层后，准确率、召回率以及 F1 值均降低，证明了自注意力层对于中文文本表示能力的必要性。

3.6 端到端的三元组抽取预测

为了进一步观察模型在知识抽取中的表现，对模型进行端到端的性能验证。即输入一个句子，然后输出该句子包含的所有三元组。其中三元组是（h，r，t）的形式，h 是主实体，t 是客实体，r 是两个实体之间的关系，predicate 代表关系预测的可能性。表 3、表 4 展示了部分科技资源文本的三元组抽取预测结果。

表 3 文本 1 的三元组抽取预测结果

Table 3 Triple extraction prediction results for Text 1

文本 1	楼天礼，男，研究员，毕业于浙江工业大学，主要从事高校信息化管理工作		
主实体 h	楼天礼		
客实体 t	浙江工业大学		
关系 r	毕业院校	出品公司	出版社
Predicate	0.885 6	0.375 9	0.323 8

表 4 文本 2 的三元组抽取预测结果

Table 4 Triple extraction prediction results for Text 2

文本 2	柑橘凤蝶西藏亚种属于动物界鳞翅目凤蝶科		
主实体 h	柑橘凤蝶西藏亚种		
客实体 t	鳞翅目		
关系 r	目	作者	成立日期
Predicate	0.962 8	0.012 5	0.009 3

从表 3 可以看出，关系是“毕业院校”的可能性最大，接近 0.9，而关系“出品公司”和“出版社”由于两者语义较为接近，所以预测可能性差别不大。从表 4 可以看出，关系“目”的可能性达到 95%以上，考虑到“目”可以算作专业词汇，同时后两者“作者”和“成立日期”由于语义相差过大，所以预测可能性几乎为零。通过以上内容，验证了模型在处理中文科技文本的知识抽取任务中的有效性。

3.7 偏置权重对模型的影响

模型引入偏置参数 α 来增强实体之间的联系。对于偏置权重参数 α，当其取值为 1 时，表示目标函数没有使用偏置损失，对于包括“O”标签所有标签都使用一样的学习权重；当其取值较大时，表示倾向于忽视“O”标签的预测结果，但也可能带来精确率下降的问题。本文通过设置参数 α 的值为 1.0，1.5，2.0，2.5，3.0，3.5，4.0，4.5，5.0，5.5，比较在不同取值情况下算法的准确率、召回率和 F1 值的变化情况，结果如图 2 所示。

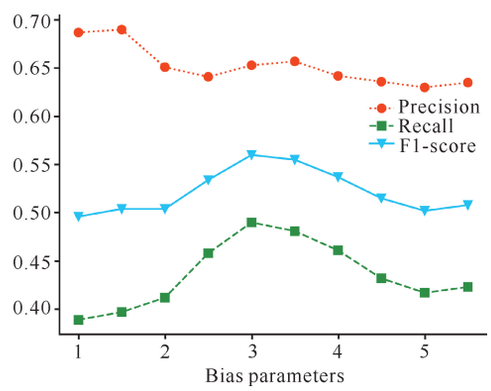

图 2 不同取值下偏置权重的模型抽取效果

Fig. 2 Model extraction effect of bias weights under different values

当 α 增大时，准确率呈逐级下降趋势；在 α 取值为 3－4 时，准确率达到最高。召回率整体呈先上升后下降的趋势，同样在 α 取值为 3 附近时，召回率达到最大值。F1 值的整体变化情况与召回率类似。综上所述，当 α 取值在 3 附近时，模型能够获得准确率和召回率之间的平衡，从而得到最高的 F1 值。因此设置偏置参数 $\alpha=3$。

4 结论

针对中文文本语义特殊性和流水式抽取方法收敛较为缓慢的问题，本文提出了一种基于字词混合和 GRU 的科技文本知识抽取（MBGAB）方法，有效提升了针对中文科技资源文本的知识抽取的效果。采用一个基于 GRU 端到端的模型来生成标柱序列，双向 GRU 对输入句子进行编码，还有一个带有偏置损失的 GRU 编码层，最后使用带有偏置权重的目标函数来增强实体标签的相关性和降低无用标签的影响度。为了在最大程度上避免边界切分出错，同时为了存储更加有效的语义信息，本文设计了一种字词混合向量映射方式。同时，结合自注意力机制，针对中文科技资源文本进行知识抽取。实验结果表明，MBGAB 在科技资源文本数据知识抽取任务中准确率、召回率以及 F1 值均有一定提升，验证了该方法的有效性。

参考文献

[1] KOU F F,DU J P,HE Y J,et al. Social network search based on semantic analysis and learning [J]. CAAI Transactions on Intelligence Technology, 2016, 1(4): 293-302.

[2] LI A,DU J P,KOU F F,et al. Scientific and technological information oriented semantics - adversarial and media- adversarial cross - media retrieval [EB/OL]. [2022-4-10]. https://arxiv.org/pdf/2203.08615v1.pdf.

[3] KOU F F,DU J P,YANG C X,et al. Hashtag recommendation based on multi-features of microblogs [J]. Journal of Computer Science and Technology, 2018, 33(4):711-726.

[4] SHI C,HAN X T,SONG L,et al. Deep collaborative filtering with multi-aspect information in heterogeneous networks [J]. IEEE Transactions on Knowledge and Data Engineering,2021,33(4):1413-1425.

[5] 杨佳鑫,杜军平,邵莹侠,等. 面向知识产权的科技资源画像构建方法[J]. 软件学报,2022,33(4):1439-1450.

[6] YOO J,CHO M,KIM T,et al. Knowledge extraction with no observable data[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: [s. n.], 2019: 2705-2714.

[7] JANG Y J,JEONG I,CHO Y K. Business failure prediction of construction contractors using a LSTM RNN with accounting, construction market, and macroeconomic variables [J]. Journal of Management in Engineering,2020,36(2):04019039. DOI:10.1061/(ASCE)ME.1943-5479.0000733.

[8] RONRAN C,LEE S,JANG H J. Delayed combination of feature embedding in bidirectional LSTM CRF for NER [J]. Applied Sciences, 2020, 10(21): 7557. DOI: 10.3390/app10217557.

[9] 温超东,曾诚,任俊伟. 结合 ALBERT 和双向门控循环单元的专利文本分类[J]. 计算机应用,2021,41(2):407-412.

[10] LIANG Z Y,DU J P,LI C Y. Abstractive social media text summarization using selective reinforced Seq2Seq attention model [J]. Neurocomputing, 2020, 410: 432-440.

[11] 刘知远,孙茂松,林衍凯,等. 知识表示学习研究进展[J]. 计算机研究与发展,2016,53(2):247-261.

[12] NADEAU D,SEKINE S. A survey of named entity recognition and classification [J]. Lingvisticae Investigationes,2007,30(1):3-26.

[13] RINK B,HARABAGIU S. Utd:Classifying semantic relations by combining lexical and semantic resources [C]//Proceedings of the 5th International Workshop on Semantic Evaluation. Uppsala, Sweden: ACL, 2010: 256-259.

[14] QIN P D,XU W R,GUO J. An empirical convolutional neural network approach for semantic relation classification [J]. Neurocomputing,2016,190:1-9.

[15] FANG Y K,DENG W H,DU J P,et al. Identity-aware CycleGAN for face photo-sketch synthesis and recognition [J]. Pattern Recognition,2020,102:107249. DOI: 10.1016/j.patcog.2020.107249.

[16] ZENG D J,LIU K,LAI S W,et al. Relation classification via convolutional deep neural network [C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: [s. n.], 2014: 2335-2344.

[17] NGUYEN T H,GRISHMAN R. Relation extraction: Perspective from convolutional neural networks [C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver, Colorado: [s. n.], 2015: 39-48.

[18] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces [C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea: ACL, 2012: 1201-1211.

[19] XU Y, MOU L, LI G, et al. Classifying relations via long short term memory networks along shortest dependency paths [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: ACL, 2015: 1785-1794.

[20] ZHANG S, ZHENG D Q, HU X C, 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: [s. n.], 2015: 73-78.

[21] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016: 1105-1116.

[22] ZHENG S C, HAO Y X, LU D Y, et al. Joint entity and relation extraction based on a hybrid neural network [J]. Neurocomputing, 2017, 257: 59-66.

[23] ZHENG S C,WANG F,BAO H Y,et al. Joint extraction of entities and relations based on a novel tagging scheme [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver,Canada:ACL,2017:1227-1236.

[24] XU L,DU J P,LI Q P. Image fusion based on nonsubsampled contourlet transform and saliency-motivated pulse coupled neural networks [J]. Mathematical Problems in Engineering, 2013: 135182. DOI: 10. 1155/2013/135182.

[25] BEKOULIS G,DELEU J,DEMEESTER T,et al. Joint entity recognition and relation extraction as a multi-head selection problem [J]. Expert Systems with Applications,2018,114:34-45.

[26] WANG S L,ZHANG Y,CHE W X,et al. Joint extraction of entities and relations based on a novel graph scheme [C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18). Vienna,Austria:IJCAI,2018:4461-4467.

[27] FU T J,MA W Y,LI P H. GraphRel:Modeling text as relational graphs for joint entity and relation extraction [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence,Italy:ACL,2019:1409-1418.

[28] MINTZ M,BILLS S,SNOW R,et al. Distant supervision for relation extraction without labeled data [C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec,Singapore:ACL,2009:1003-1011.

Knowledge Extraction Method of Scientific and Technological Text Based on Word Mixing and GRU

OUYANG Suyu,SHAO Yingxia,DU Junping,LI Ang

(Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,College of Computer Science,Beijing University of Posts and Telecommunicates,Beijing,100082,China)

Abstract:The knowledge extraction task is to extract triple relations (head entity-relation-tail entity) from the unstructured text data. The existing knowledge extraction methods are divided into "pipeline" method and joint extraction method. The "pipeline" method extracts named entity recognition and entity knowledge extraction with their respective modules. Although this method has better flexibility,the training speed is slow. The learning model of joint extraction is an end-to-end model implemented by neural network to realize entity recognition and relationship extraction at the same time,which can well preserve the association between entities and relationships,and convert the joint extraction of entities and relationships into a sequence labeling problem. The main contributions of this paper are as follows:① A knowledge extraction method for scientific and technological text based on word mixing and Gated Recurrent Unit (MBGAB) is proposed,which combines attention mechanism to extract the relationship between Chinese scientific and technological resource text. ② Vector mapping method using mixed words can not only avoid boundary segmentation errors to the greatest extent,but also effectively integrate semantic information. ③ The end-to-end joint extraction model,the bidirectional GRU network and the self-attention mechanism are used to effectively capture the long-distance semantic information in the sentence,and the bias weight is introduced to improve the effect of model extraction.

Key words:knowledge extraction;vector map;GRU;triple relation;joint extraction method

责任编辑:陆媛峰