DOI:10.7544/issn1000-1239.2020.20190748
57(11):2456-2466,2020

基于类卷积交互式注意力机制的属性抽取研究

尉桢楷　程　梦　周夏冰　李志峰　邹博伟　洪　宇　姚建民
(苏州大学计算机科学与技术学院　江苏苏州　215006)
(20185227064@stu.suda.edu.cn)

Convolutional Interactive Attention Mechanism for Aspect Extraction

Wei Zhenkai, Cheng Meng, Zhou Xiabing, Li Zhifeng, Zou Bowei, Hong Yu, and Yao Jianmin
(*College of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

Abstract Attention mechanism is a common model in aspect extraction research. There are two limitations in attention mechanism towards aspect extraction: First, existing attention mechanism is mostly static attention or self attention.Self attention mechanism is a global attention mechanism, and it brings the irrelevant noises (words that are far away from the target word and unrelated to it) into attention vector; Second, existing attention mechanisms are mostly single-layer which lack interactivity. To address above two limitations, a convolutional interactive attention (CIA) mechanism is proposed in this paper. A bidirectional long short term memory network (Bi-LSTM) is exploited to obtain hidden representations of words in a target sentence, and then the convolutional interactive attention mechanism is used for representation learning. Convolutional interactive attention mechanism includes two layers: in the first layer, the number of context words for each target word is limited by a window, then the context words are used to calculate the attention vector of target word. In the second layer, the interactive attention vector is calculated by attention distribution of the first layer and all the words in target sentence. After that, we concatenate attention vectors of the first layer and second layer. Finally, conditional random field (CRF) is utilized to label aspects. This paper demonstrates the effectiveness of the proposed method over the official evaluation datasets of 2014—2016 Semantic Evaluation (SemEval).Compared with the baseline, the model proposed in this paper increases the $F1$ score of aspect extraction with 2.21%, 1.35%, 2.22% and 2.21% respectively on four datasets.

Key words attention mechanism; aspect extraction; conditional random field (CRF); sequence labeling; sentiment analysis

摘　要 在基于深度学习的属性抽取研究中,注意力机制是常用的模型之一.目前,面向属性抽取的注意力机制存在2个局限性:其一,注意力机制多为自注意力机制,这是一种全局式注意力机制,其将不相关的噪音(距离目标词较远且与之不相关的词)带入注意力向量的计算;其二,目前的注意力机制多为单层注意力机制,注意力一次建模后缺少交互性.针对这2个局限性,提出一种面向属性抽取的类卷积交互式注意力机制.该方法先将目标句输入到双向循环神经网络,借以获得每个词的隐式表达,再经过类卷积交互式注意力机制进行表示学习.类卷积交互式注意力机制分为2层注意力计算:第1层按序(从句首到句末)通过滑动窗口控制每个词的上下文宽度,并计算每个词的注意力分布向量;第2层将第1层

收稿日期:2019-10-21;**修回日期**:2020-02-13
基金项目:国家自然科学基金项目(61672367,61672368,61703293)
This work was supported by the National Natural Science Foundation of China (61672367, 61672368, 61703293).
通信作者:周夏冰(zhouxiabing@suda.edu.cn)

的注意力分布向量与所有单词进行交互注意力计算,将得到的注意力向量与第1层的注意力向量拼接,最终输入到条件随机场进行属性标记.在2014—2016语义评估(semantic evaluation, SemEval)官方数据集上验证了模型的有效性.相比于基线模型,在4个数据集上的$F1$值分别提高了2.21,1.35,2.22,2.21个百分点.

关键词 注意力机制;属性抽取;条件随机场;序列标注;情感分析

中图法分类号 TP391

属性抽取(aspect extraction)是属性级情感分析的子任务之一[1],其目标是:对于用户评价的文本,抽取其中用户所评价的属性或实体.表1给出了3条评价文本样例,前2条为餐馆领域评价文本,其中"cheesecake(奶酪蛋糕)"、"pastries(糕点)"、"food(食物)"、"dishes(菜肴)"为待抽取的属性,粗体显示;最后一条为电脑领域评价文本,其中待抽取的属性为"screen(屏幕)"、"clicking buttons(点击按钮)",粗体表示.

Table 1 Example of User Review

表1 评价文本样例

No.	Sentences
1	I got an excellent piece of **cheesecake** and we had several other nice **pastries.**
2	The **food** was average or above including some surprising tasty **dishes.**
3	The **screen** is a little glary, and I hated the **clicking buttons**, but I got used to them.

目前,针对属性抽取的研究方法主要分为3类:基于规则的方法、基于传统机器学习的方法和基于深度学习的方法.基于规则的方法依赖于领域专家制定的规则模板实现属性抽取.例如,Hu等人[2]首次提出使用关联规则实现属性抽取,并且只抽取评论文本中显式的名词属性或名词短语属性.Li等人[3]使用依存关系从影评中抽取"评价对象-评价意见"单元对.Qiu等人[4]利用依存关系获得属性词与评价词之间的关系模板,从而根据属性词抽取评价词,根据评价词抽取属性词.以上基于规则的方法迁移性差,无法抽取规则之外的属性.在基于传统机器学习的方法中,通常将属性抽取任务指定为序列标注任务.其中,Jakob等人[5]首次将条件随机场(conditional random field, CRF)应用于属性抽取的研究,并融合了多种特征,在属性抽取的任务上取得了较好的效果.Xu等人[6]在CRF的基础上引入浅层句法分析和启发式位置特征,在不增加领域词典的情况下,有效地提高了属性抽取的性能.然而,基于CRF的模型通常依赖于大量的手工特征,在特征缺失的情况下性能将会大幅下降.

深度学习的方法可以避免大量的手工特征,自动学习特征的层次结构完成复杂的任务,在属性抽取的任务上取得了优异的效果.例如,Liu等人[7]首次将长短期记忆网络(long-short term memory, LSTM)应用于属性抽取任务,与使用大量手工特征的CRF模型相比,该方法取得了更优的性能.Toh等人[8]提出将双向循环神经网络(bidirectional recurrent neural network, Bi-RNN)与CRF相结合的方法,在2016年SemEval属性级情感分析评测任务中性能达到最优.

目前,注意力机制(attention mechanism)已被应用于属性抽取的研究.Wang等人[9]提出一种多任务注意力模型,将属性词和情感词的抽取与分类进行联合训练,从而实现学习抽取和分类过程中的特征共享,进而实现抽取和分类的相互促进,该模型应用的注意力机制为静态注意力机制.Cheng等人[10]在基于双向长短期记忆网络的CRF模型(BiLSTM-CRF)中着重利用门控动态注意力机制,所使用的注意力机制为自注意力机制.BiLSTM-CRF的架构[11-13]既捕获了句子中上下文的分布特征,又有效地利用上下文标记预测当前的标记类别,鉴于此本文将BiLSTM-CRF的架构作为基线模型.

目前面向属性抽取的注意力机制存在2个局限性.其一,注意力机制多为全局式注意力机制(本文将自注意力机制统称为全局式注意力机制),全局式注意力机制在每个时刻(处理每个目标词项时)将与之距离较远且关联不密切的词分配了注意力权重.例如,评论句子"The service is great, but the icecream is terrible."(译文:服务很好,但冰淇淋糟糕),当目标词为"service(服务)"时,"terrible(糟糕)"距离目标词"service"较远且关联不紧密,若对"terrible"分配较高的注力权重,则为目标词"service"的注意力分布向量带来噪音.其二,目前面向属性抽取的注意力机制多为单层,注意力机制单层建模后缺少交互性.

针对上述局限,本文提出面向属性抽取的类卷积交互式注意力机制(convolutional interactive attention, CIA).该注意力机制在每个时刻(处理每个目标词时)都通过滑动窗口控制目标词的上下文词的个数,例如图 1,当前时刻的目标词为"icecream(冰淇淋)"时,在滑动窗口内计算"icecream"的注意力分布向量.在此基础上,再将目标词的注意力分布向量与句中各个词进行交互注意力计算,将获得的交互注意力向量与目标词的注意力分布向量拼接,由此获得最终的注意力分布向量.

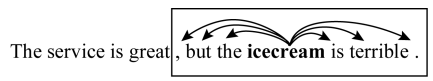

Fig. 1 Example of attention vector computation of target word in a window

图 1 窗口内目标词的注意力向量计算样例

本文提出在 BiLSTM-CRF 的基础上着重利用 CIA 的模型 CIA-CRF,CIA-CRF 是针对属性抽取任务形成的一种综合神经网络和 CRF 的架构,在该架构中配以一套新型的注意力机制 CIA.总体上,本文的贡献包含 2 个方面:

1) 提出类卷积交互式注意力机制(即 CIA),该注意力机制分为类卷积注意力层和交互注意力层,旨在解决目前面向属性抽取的全局式注意力机制将不相关的噪音带入注意力向量的计算以及注意力机制缺少交互性的局限.

2) 利用 Bi-LSTM 对句中所有的词提取字符级特征,将字符级特征与各自的词向量拼接,以此获得含有字符级特征的词向量表示.字符级特征有助于未登录词的识别.

本文在国际属性级情感分析公开数据集 SemEval 2014[1], 2015[14], 2016[15] 上对 CIA-CRF 进行测试,在 4 个数据集上 $F1$ 值均获得提升.

1 模 型

1.1 属性抽取任务

与 Yu 等人[16]方法类似,本文将属性抽取任务指定为序列标注任务,使用的标签模式为 BMESO.对于包含多个词的属性,B 代表属性的开端,M 代表属性的中间,E 代表属性的结尾;对于单个词的属性,则用 S 表示;O 统一代表非属性词.序列标注样例如表 2 所示:

Table 2 Example of Sequence Labeling

表 2 序列标注样例

Sequence	The	food	portion	sizes	are	appropriate
Label	O	B	M	E	O	O

1.2 模型总体结构

本文采用 BiLSTM-CRF 的模型架构,并结合了类卷积交互式注意力机制,该注意力机制分为类卷积注意力层和交互注意力层.本文方法的总体结构如图 2 所示.首先对于一个待抽取句子 $X=\{x_1, x_2, \cdots, x_n\}$,初始化每个词的分布式表示,本文采用预训练词向量 $\boldsymbol{e}_i (i=1,2,\cdots,n)$ 作为词义的分布式表示,矩阵 $\boldsymbol{E}=(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n)$.由于每个词都由字符组成,因此可将各个词转化为字符矩阵 $(\boldsymbol{e}_i^1, \boldsymbol{e}_i^2, \cdots, \boldsymbol{e}_i^{L_i})$ 表示,其中 $L_i (i=1,2,\cdots,n)$ 为词的字符个数.矩阵 $\boldsymbol{E}'=((\boldsymbol{e}_1^1, \boldsymbol{e}_1^2, \cdots, \boldsymbol{e}_1^{L_1}), (\boldsymbol{e}_2^1, \boldsymbol{e}_2^2, \cdots, \boldsymbol{e}_2^{L_2}), \cdots, (\boldsymbol{e}_n^1, \boldsymbol{e}_n^2, \cdots, \boldsymbol{e}_n^{L_n}))$.在此基础上,概述模型的各层功能.

1) 将词 x_i 的字符矩阵 $(\boldsymbol{e}_i^1, \boldsymbol{e}_i^2, \cdots, \boldsymbol{e}_i^{L_i})$ 输入到 Bi-LSTM 中进行编码(以第 i 个词为例),取最后时刻的隐藏状态与词向量 $\boldsymbol{e}_i$ 拼接,从而获得拼接向量 $\boldsymbol{s}_i$,可获得各个词含有字符级特征的分布式表示 $\boldsymbol{S}=(\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n)$;

2) 将 $\boldsymbol{S}=(\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n)$ 输入 Bi-LSTM 层,通过 Bi-LSTM 的编码,借以获得各个词包含上下文信息的隐藏状态 $\boldsymbol{H}=(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n)$;

3) 将 $\boldsymbol{H}=(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n)$ 经过类卷积注意力层,按序(从句首到句尾)逐词地对距离各个词较近的若干词分配注意力权重,进而通过注意力权重及其对应的隐藏状态计算类卷积注意力矩阵 $\boldsymbol{H}'=(\boldsymbol{h}_1', \boldsymbol{h}_2', \cdots, \boldsymbol{h}_n')$;

4) 将 $\boldsymbol{H}'$ 经过交互注意力层,按序逐词地对各个单词的上下文所有词分配注意力权重,进而通过注意力权重和类卷积注意力矩阵 $\boldsymbol{H}'$ 计算交互注意力矩阵 $\boldsymbol{Q}=(\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n)$,最后将类卷积注意力矩阵 $\boldsymbol{H}'$ 与交互注意力矩阵 $\boldsymbol{Q}$ 拼接,由此获得双层注意力矩阵表示 $\boldsymbol{R}=(\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_n)$;

5) 经过注意力层的表示学习后,本文继承 Cheng 等人[10]的工作,将双层注意力矩阵 $\boldsymbol{R}$ 输入到门控循环单元(gated recurrent unit, GRU)中更新,从而获得更新后的注意力矩阵 $\boldsymbol{U}=(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n)$,并经过全连接降维后输入到 CRF 层进行属性标记,最终获取各个单词对应的预测标签 $L=\{l_1, l_2, \cdots, l_n\}$,其中 $l_i \in \{\mathrm{B}, \mathrm{M}, \mathrm{E}, \mathrm{S}, \mathrm{O}\}$.

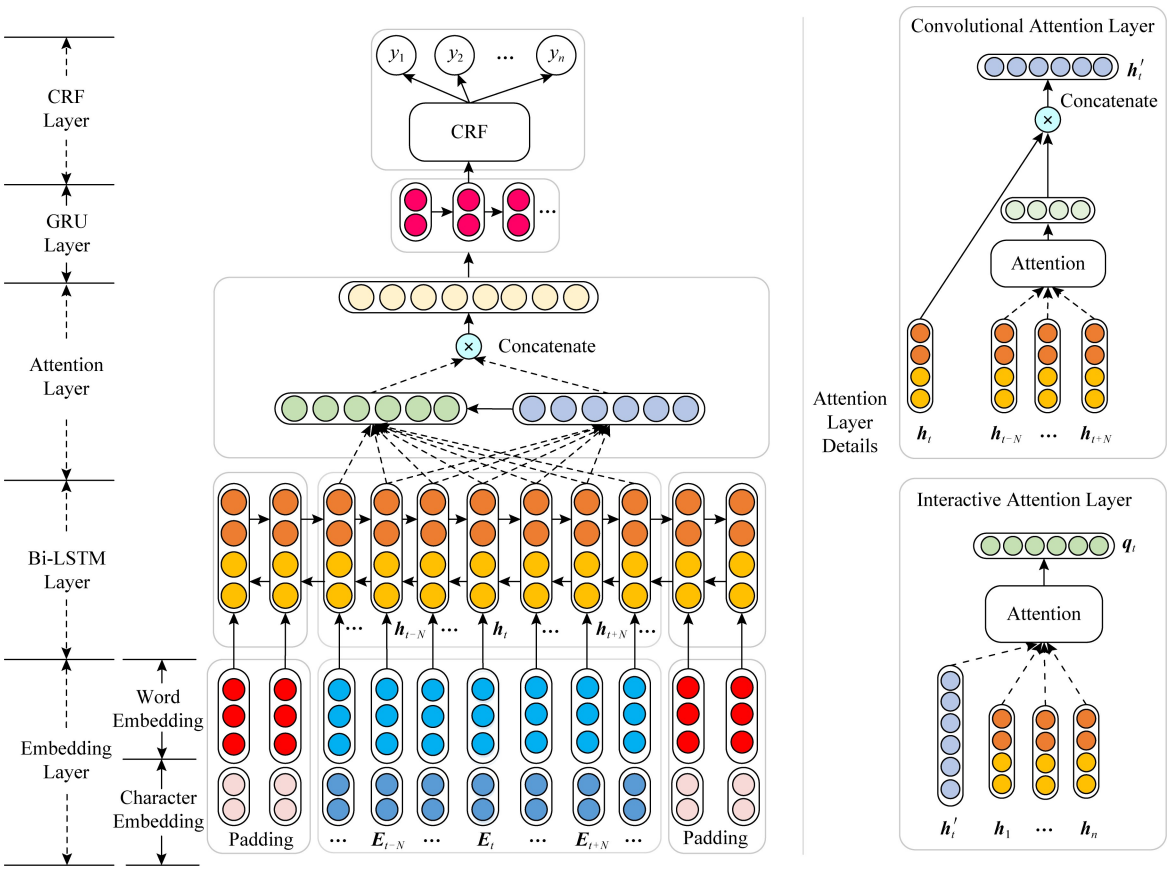

Fig. 2 General structure of system

图 2 模型总体结构

1.3 词语表示层

本文继承 Lample 等人[17]在命名实体识别研究中的工作,将当前时刻词 x_t(当前词)的字符矩阵 $(\boldsymbol{e}_t^1,\boldsymbol{e}_t^2,\cdots,\boldsymbol{e}_t^{L_t})$ 输入到 Bi-LSTM 进行编码,取 Bi-LSTM 最后时刻的隐藏状态与当前词的词向量 $\boldsymbol{e}_t$ 拼接,从而获得含有字符级特征的词向量 $\boldsymbol{s}_t$.词语表示层的结构如图 3 所示,具体的计算公式为:

$$\overrightarrow{\boldsymbol{h}}_t^L = lstm(\overrightarrow{\boldsymbol{h}}_t^{L-1},\boldsymbol{e}_t^L,\overrightarrow{\boldsymbol{\theta}}), \tag{1}$$

$$\overleftarrow{\boldsymbol{h}}_t^1 = lstm(\overleftarrow{\boldsymbol{h}}_t^2,\boldsymbol{e}_t^1,\overleftarrow{\boldsymbol{\theta}}), \tag{2}$$

$$\boldsymbol{s}_t = [\overrightarrow{\boldsymbol{h}}_t^L;\overleftarrow{\boldsymbol{h}}_t^1;\boldsymbol{e}_t], \tag{3}$$

式中,$lstm$ 为 LSTM 模型,$\overrightarrow{\boldsymbol{h}}_t^L$ 为正向 $lstm$ 最后时刻的隐藏状态,$\overleftarrow{\boldsymbol{h}}_t^1$ 为反向 $lstm$ 最后时刻的隐藏状态,$\boldsymbol{e}_t$ 为当前词的词向量,$\overrightarrow{\boldsymbol{\theta}}$ 和 $\overleftarrow{\boldsymbol{\theta}}$ 为 $lstm$ 的参数矩阵.

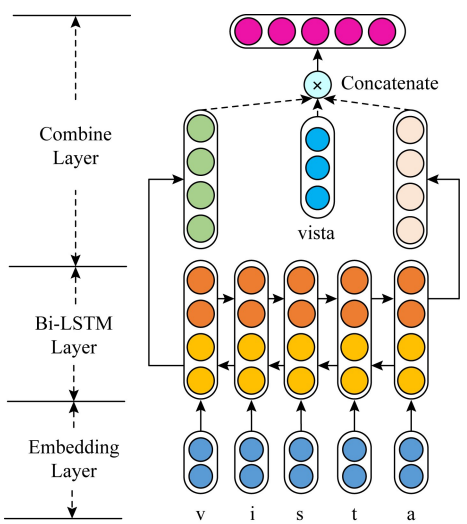

Fig. 3 Structure of word representation layer

图 3 词语表示层结构

1.4 Bi-LSTM 层

由 1.3 节可以获得各个含有字符特征的词矩阵 $\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n)$,本文采用 Bi-LSTM 对词矩阵 $\boldsymbol{S}$ 进行编码.

Bi-LSTM 由前向 LSTM 和后向 LSTM 组合而成.其中,LSTM 有 3 个输入,分别是当前时刻的输入 $\boldsymbol{s}_t$、上一时刻 LSTM 的输出 $\boldsymbol{h}_{t-1}$、上一时刻的记忆单元状态 $\boldsymbol{c}_{t-1}$,LSTM 的输出有 2 个,分别是当前时刻的输出 $\boldsymbol{h}_t$ 和当前时刻的记忆单元状态 $\boldsymbol{c}_t$.LSTM 的内部结构由 3 个门组成,依次为遗忘门

$\boldsymbol{f}_t$、输入门 $\boldsymbol{i}_t$、输出门 $\boldsymbol{o}_t$.3 个门控的功能各不相同,遗忘门选择通过的信息量,输入门控制当前输入对记忆单元状态的影响,输出门控制输出信息.LSTM 的计算公式为:

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\mathrm{sf}}\boldsymbol{s}_t+\boldsymbol{W}_{\mathrm{hf}}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f),\tag{4}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{\mathrm{si}}\boldsymbol{s}_t+\boldsymbol{W}_{\mathrm{hi}}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i),\tag{5}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{\mathrm{so}}\boldsymbol{s}_t+\boldsymbol{W}_{\mathrm{ho}}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o),\tag{6}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tanh(\boldsymbol{W}_{\mathrm{sc}}\boldsymbol{s}_t+\boldsymbol{W}_{\mathrm{hc}}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c),\tag{7}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t),\tag{8}$$

式中 σ 为 sigmod 激活函数,tanh 为 tanhyperbolic 激活函数;$\boldsymbol{W}$ 表示权重矩阵,$\boldsymbol{b}$ 表示偏置项.

传统的 LSTM 只能捕捉到正向的语义信息,但无法捕捉到未来的上下文信息.因此,本文使用 Bi-LSTM 模型,正向 LSTM 捕捉当前词的上文信息,获得正向特征 $\overrightarrow{\boldsymbol{h}}_t$,反向的 LSTM 捕捉当前词的下文信息,获得反向特征 $\overleftarrow{\boldsymbol{h}}_t$.Bi-LSTM 通过拼接正反向特征,以此获得当前词的隐层表示 $\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t;\overleftarrow{\boldsymbol{h}}_t]$.

1.5 类卷积交互式注意力机制

本文针对属性抽取任务,提出一种面向属性抽取的类卷积交互式注意力机制方法.该注意力机制为双层注意力机制.第 1 层为类卷积注意力层,旨在降低全局式注意力机制在计算注意力向量时带入的噪声;第 2 层为交互注意力层,是在类卷积注意力层降噪的基础上引入的.之所以提出交互注意力层,是由于在类卷积注意力层中,滑动窗口大小为固定的超参数,所以窗口外可能存在与当前词关联密切的词.基于类卷积注意力向量,与所有词做进一步地交互注意力计算,从而获得对于类卷积注意力向量而言重要的全局信息.因此,类卷积交互式注意力机制既满足了降噪,又获得对于类卷积注意力向量而言重要的全局信息.

总之,类卷积注意力层布置于交互注意力层之前,专用于去噪.从而再次使用交互注意力层时,噪声已获得类卷积注意力层的处理,同时保留了交互注意力层自身的优势.下面将分别详细介绍类卷积注意力层和交互注意力层.

1.5.1 类卷积注意力层

Kim[18]首次将卷积神经网络应用于文本分类任务,通过卷积核获取每个目标词的上下文特征.我们将这种卷积思想迁移到注意力机制的计算,设置类似于卷积核的滑动窗口,通过滑动窗口的大小限制每个目标词的上下文词的个数,从而在滑动窗口内计算每个目标词的类卷积注意力向量.类卷积注意力层如图 4 所示:

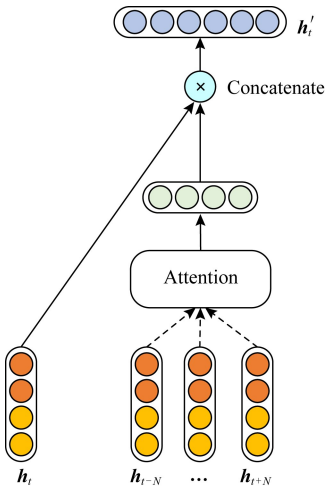

Fig. 4 Convolutional attention layer

图 4 类卷积注意力层

由 1.4 节,我们可以得到各个词的隐藏层表示 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$,$\boldsymbol{h}_t$ 为当前时刻 t(即第 t 个词)的隐藏状态,$\boldsymbol{h}_i^t(i\in[t-N,t-1]\cup[t+1,t+N])$为窗口内目标词 $\boldsymbol{h}_t$ 的上下文隐藏状态,其中 N 为目标词的上文(下文)词的个数.经过式(9)的计算:

$$a_i^t=\boldsymbol{V}^{\mathrm{T}}\tanh(\boldsymbol{W}_1\boldsymbol{h}_i^t+\boldsymbol{W}_2\boldsymbol{h}_t),\tag{9}$$

其中,$\boldsymbol{W}_1$ 与 $\boldsymbol{W}_2$ 是窗口内目标词及其上下文的权重矩阵,$\boldsymbol{V}$ 是调节非归一化注意力权重的参数矩阵.可获得窗口内目标词 $\boldsymbol{h}_t$ 的上下文的非归一化注意力得分 $\boldsymbol{A}=(a_{t-N}^t,\cdots,a_{t-1}^t,a_{t+1}^t,\cdots,a_{t+N}^t)$,$\boldsymbol{A}$ 表示窗口内目标词的上下文注意力权重的分配.为了使注意力权重在统一的标准中进行公平计算,在式(10)中,对非归一化注意力得分 $\boldsymbol{A}$ 进行归一化处理,进而获得归一化后的注意力得分 $\boldsymbol{K}=(k_{t-N}^t,\cdots,k_{t-1}^t,k_{t+1}^t,\cdots,k_{t+N}^t)$.

$$k_i^t=\frac{\exp(a_i^t)}{\sum_{j=t-N}^{t-1}\exp(a_j^t)+\sum_{j=t+1}^{t+N}\exp(a_j^t)}.\tag{10}$$

本文将归一化后的注意力得分 $\boldsymbol{K}=(k_{t-N}^t,\cdots,k_{t-1}^t,k_{t+1}^t,\cdots,k_{t+N}^t)$与窗口内对应的隐藏状态 $\boldsymbol{h}_{t-N},\cdots,\boldsymbol{h}_{t-1},\boldsymbol{h}_{t+1},\cdots,\boldsymbol{h}_{t+N}$ 加权求和,从而得到目标词的类卷积注意力向量 $\boldsymbol{h}'_t$,并将类卷积注意力向量 $\boldsymbol{h}'_t$ 与对应的隐藏状态 $\boldsymbol{h}_t$ 拼接,计算为:

$$\boldsymbol{h}'_t=\sum_{i=t-N}^{t-1}(k_i^t\times\boldsymbol{h}_i^t)+\sum_{i=t+1}^{t+N}(k_i^t\times\boldsymbol{h}_i^t),\tag{11}$$

$$\boldsymbol{h}'_t=[\boldsymbol{h}'_t;\boldsymbol{h}_t].\tag{12}$$

1.5.2 交互注意力层

交互注意力层如图 5 所示.第 t 个词的类卷积注意力向量 $\boldsymbol{h}'_t$ 分别与各个词的隐藏状态 $\boldsymbol{H}=(\boldsymbol{h}_1,$

$\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n$)相加,从而获得基于类卷积注意力向量的隐藏状态 $\boldsymbol{D}^t=(\boldsymbol{d}_1^t,\boldsymbol{d}_2^t,\cdots,\boldsymbol{d}_n^t)$,公式为:

$$\boldsymbol{d}_j^t=\boldsymbol{h}_t'+\boldsymbol{h}_j, \tag{13}$$

其中 $\boldsymbol{h}_j$ 为第 j 个词的隐藏状态,$j\in[1,n]$.

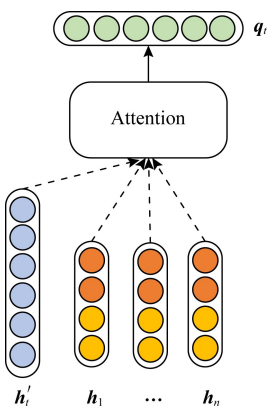

Fig. 5 Interactive attention layer

图 5 交互注意力层

在当前时刻 t,通过类卷积注意力向量 $\boldsymbol{h}_t'$ 和基于类卷积注意力向量的隐藏状态 $\boldsymbol{D}^t=(\boldsymbol{d}_1^t,\boldsymbol{d}_2^t,\cdots,\boldsymbol{d}_n^t)$,经过式(14)的计算:

$$g_j^t=\boldsymbol{M}^{\mathrm{T}}\tanh(\boldsymbol{W}_3\boldsymbol{d}_j^t+\boldsymbol{W}_4\boldsymbol{h}_t'+\boldsymbol{b}), \tag{14}$$

其中,$\boldsymbol{W}_3,\boldsymbol{W}_4$ 分别为类卷积注意力向量和基于类卷积注意力向量的隐藏状态的权值矩阵,$\boldsymbol{b}$ 为偏置矩阵,$\boldsymbol{M}$ 为调节非归一化交互注意力权重的参数矩阵.获得 $\boldsymbol{d}_1^t,\boldsymbol{d}_2^t,\cdots,\boldsymbol{d}_n^t$ 的非归一化交互注意力权重 $\boldsymbol{G}=(g_1^t,g_2^t,\cdots,g_n^t)$.为了使权重在统一的标准下公平计算,本文对非归一化注意力得分 $\boldsymbol{G}$ 进行归一化处理,以此获得在统一标准下衡量的注意力得分 $\boldsymbol{C}=(c_1^t,c_2^t,\cdots,c_n^t)$,归一化公式为:

$$c_j^t=\frac{\exp(g_j^t)}{\sum_{m=1}^{n}\exp(g_m^t)}. \tag{15}$$

为了获得注意力分布不同的交互注意力向量,本文将归一化后的交互注意力得分 $\boldsymbol{C}=(c_1^t,c_2^t,\cdots,c_n^t)$与对应的基于类卷积注意力向量的隐藏状态 $\boldsymbol{D}^t=(\boldsymbol{d}_1^t,\boldsymbol{d}_2^t,\cdots,\boldsymbol{d}_n^t)$加权求和,从而得到交互注意力向量 $\boldsymbol{q}_t$.计算公式为:

$$\boldsymbol{q}_t=\sum_{j=1}^{n}c_j^t\times\boldsymbol{d}_j^t. \tag{16}$$

本文将目标词的类卷积注意力向量 $\boldsymbol{h}_t'$ 与交互注意力向量 $\boldsymbol{q}_t$ 拼接,从而获得目标词的双层注意力向量表示 $\boldsymbol{r}_t$,进而将句中各个目标词的双层注意力向量表示为矩阵 $\boldsymbol{R}=(\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_n)$.本文继承 Cheng 等人[10]的工作,将双层注意力矩阵 $\boldsymbol{R}=(\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_n)$输入到 GRU 网络中,借以获取更新后的各个词的注意力向量,表示为矩阵 $\boldsymbol{U}=(\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_n)$,计算公式为:

$$\boldsymbol{r}_t=[\boldsymbol{h}_t';\boldsymbol{q}_t], \tag{17}$$

$$\boldsymbol{u}_t=gru(\boldsymbol{u}_{t-1},\boldsymbol{r}_t,\boldsymbol{\theta}), \tag{18}$$

其中,gru 为 GRU 模型,$\boldsymbol{\theta}$ 为 gru 的参数矩阵.

1.6 CRF

CRF 最早由 Lafferty 等人[19]于 2001 年提出,是一种判别式模型.线性链条件随机场被广泛应用于序列标注任务,其优越性已被多次证明.CRF 的主要作用是进一步增强前后标签的约束,避免不合法标签的出现,例如标签 M 的前一个标签是 O,即为不合法标签,CRF 输出的是合法并且概率最大的标签组合.CRF 原理为:

$$p(Y|\boldsymbol{U})=\frac{1}{Z(\boldsymbol{U})}\exp\Big(\sum_{i=1}^{n+1}\sum_{k=1}^{T}\lambda_k t_k(y_{i-1},y_i,\boldsymbol{u},i)+\sum_{i=1}^{n}\sum_{l=1}^{S}\mu_l s_l(y_i,\boldsymbol{u},i)\Big), \tag{19}$$

其中,T 是转移特征函数的数量,S 是状态特征函数的个数,$\boldsymbol{u}$ 为降维后的类卷积交互式注意力向量,Y 为输出标签,$p(Y|\boldsymbol{U})$表示在输入为 $\boldsymbol{U}$ 的情况下标签为 Y 的概率,$Z(\boldsymbol{U})$是归一化因子.$t_k(y_{i-1},y_i,\boldsymbol{u},i)$为转移特征函数,其依赖于当前位置 y_i 和前一位置 y_{i-1},λ_k 是转移特征函数对应的权值.s_l 为状态特征函数,依赖于当前位置 y_i,μ_l 是状态特征函数对应的权值.特征函数的取值为 1 或 0,以转移特征函数为例,当 $y_{i-1},y_i,\boldsymbol{u}$ 满足转移特征函数时,则特征函数取值为 1,否则取值为 0.状态特征函数同样如此.

在训练 CRF 时,使用极大似然估计的方法训练模型中的各个变量,对于训练数据$(\boldsymbol{U},Y)$,优化函数为:

$$Loss=-\sum_{i=1}^{n}\ln P(y_i|\boldsymbol{u}_i). \tag{20}$$

经过训练使得 $Loss$ 最小化.测试时,选取概率最大的一组标签序列作为最终的标注结果.

2 实 验

2.1 实验语料与实验设置

本文的实验数据来自 SemEval 2014—2016 属性级情感分析的 4 个基准数据集,数据集分为电脑(laptop)领域和餐馆(restaurant)领域.4 个基准数据集分别为:2014 年语义评测任务 4 中的电脑领域(SemEval 2014 task 4 laptop, L-14)、2014 年语义

评测任务 4 中的餐馆领域(SemEval 2014 task 4 restaurant, R-14)、2015 年语义评测任务 12 中的餐馆领域(SemEval 2015 task 12 restaurant, R-15)、2016 年语义评测任务 5 中的餐馆领域(SemEval 2016 task 5 restaurant, R-16).实验过程中,随机从训练数据中选取 20%的样本作为开发集.各个数据集的训练集、开发集以及测试集的样本数量如表 3 所示.此外,表 3 还统计了各个数据集训练样本的平均长度.

Table 3 Statistics of Datasets

表 3 语料统计

Dataset	#Train-set	#Dev-set	#Test-set	Average Length
L-14	2 436	609	800	16.72
R-14	2 433	608	800	15.38
R-15	1 052	263	685	13.99
R-16	1 600	400	676	14.37

本文使用的预训练词向量的来源为 Glove,词向量的维度为 100 维,将词的隐含变量(hidden size)以及更新注意力的 GRU 神经网络隐含变量(GRU size)同设为 100 维,字符的隐含变量(character size)、注意力向量维度(attention size)分别设为 20 和 200,学习率(learning rate)的大小设为 0.001,批量大小(batch size)设为 20,各个目标词项的上文(下文)词的个数(N)设为 5.为了防止过拟合,在各层间加入 $dropout$,设 $dropout=0.5$.梯度优化使用 adam 优化器.

2.2 评价标准

与 Yu 等人[16]相同,本文采用 $F1$ 值作为评价标准,评价过程采用精确匹配,只有当模型预测的结果与正确答案完全匹配才看作正确预测答案,换言之,预测答案从起始位置到结束位置的各个词必须与正确答案的各个词对应相同.例如,真实的答案为"sardines with biscuits",如果模型预测的答案是"biscuits",则不是正确答案.

2.3 实验对比模型

为了验证本文提出模型的有效性,本文设置 3 组对比模型.

第 1 组对比模型为传统的融入大量手工特征的模型,具体模型为:

1) HIS-RD, DLIREC, EliXa.分别为 L-14, R-14, R-15 属性抽取排名第一的评测模型.其中 HIS-RD[20]与 DLIREC[21]基于 CRF, EliXa[22]基于隐马尔可夫模型,并且它们都使用了大量的手工特征.

2) CRF.融合基本特征以及 Glove 词向量[23]的 CRF 模型.

第 2 组对比模型是将深度学习的方法应用于属性抽取任务,对比模型为:

1) LSTM.Liu 等人[7]使用 LSTM 对词向量编码,并通过最后一层全连接获得每个词的概率分布.

2) DTBCSNN+F.Ye 等人[24]提出基于依存树的卷积堆栈神经网络的方法,该方法提取的句法特征用于属性抽取.

3) MIN.Li 等人[25]提出一种基于 LSTM 的联合学习模型,使用 2 个 LSTM 联合抽取属性词和评价词,使用第 3 个 LSTM 判别情感句和非情感句.

4) MTCA.Wang 等人[9]提出一种多任务注意模型,该模型是属性抽取和属性分类的联合学习模型.

5) GMT.Yu 等人[16]提出基于多任务神经网络全局推理的模型,该模型联合抽取属性词和评价词.

第 3 组对比模型是本文的基线模型以及在基线模型基础上引入全局式注意力机制:

1) BiLSTM+CRF.在 Toh 等人[8]提出的基 Bi-RNN 的 CRF 模型上,将 Bi-RNN 替换为 Bi-LSTM.本文将 BiLSTM+CRF 作为基线模型.

2) GA-CRF.在 BiLSTM+CRF 模型的基础上,以一种全局式注意力的计算方式,对 Bi-LSTM 的输出进行全局式注意力计算.

3) CA-CRF.在 BiLSTM+CRF 模型的基础上,集成本文提出的类卷积注意力层.

4) CIA-CRF.在 BiLSTM+CRF 基础上,集成本文提出的类卷积交互式注意力机制和字符级特征.

2.4 实验结果与分析

本文提出的模型以及对比模型的实验结果如表 4 所示.从表 4 中可知,本文的模型 CIA-CRF 在 L-14, R-14, R-16 数据集上取得了最优的 $F1$ 值.

本文将 CIA-CRF 与现有方法进行比较分析.为了验证类卷积注意力层的有效性,本文在基线模型的基础上分别引入全局式注意力机制和类卷积注意力层,并进行比较分析.由于类卷积注意力层中的滑动窗口大小是重要超参数,所以本文比较分析滑动窗口大小对实验性能的影响.随后分别分析交互注意力层的有效性和字符级特征的有效性.将预训练模型 BERT(bidirectional encoder representations from transformers)[26]分别与基线模型以及引入类卷积交互式注意力机制的基线模型进行结合,从而在结合 BERT 的前提下验证类卷积交互式注意力机制的有效性.

Table 4 $F1$ Performance Comparison

表 4 模型性能($F1$)对比 %

Model	L-14	R-14	R-15	R-16
HIS-RD[20]	74.55	79.62		
DLIREC[21]	73.78	84.01		
EliXa[22]			70.05	
CRF	74.01	82.33	67.54	69.56
LSTM[7]	75.00	82.06	64.30	71.26
DTBCSNN+F[24]	75.66	83.97		
MIN[25]	77.58			73.44
MTCA[9]	69.14		71.31	73.26
GMT[16]	78.69	84.50	70.53	
BiLSTM+CRF	76.91	83.56	68.29	71.41
GA-CRF	77.40	83.74	69.00	72.20
CA-CRF	77.90	84.57	69.22	72.81
CIA-CRF	79.12	84.91	70.51	73.62

2.4.1 与现有传统模型和深度学习模型比较

在表 4 中,本文将 CIA-CRF 与现有传统模型和深度学习模型进行了比较.与融入多种手工特征的传统模型(HIS-RD,DLIREC,EliXa,CRF)相比,本文的模型 CIA-CRF 在 L-14,R-14,R-15 数据集上均取得了最优的性能并且优势明显.传统模型(HIS-RD,DLIREC,EliXa,CRF)都使用将近 10 种不同的手工特征,然而在 Bi-LSTM 结合 CRF 的架构下引入本文提出的类卷积交互式注意力机制和字符级特征,取得了比融入大量手工特征的传统模型更优越的性能.

对近年来的深度学习模型进行比较分析.相比于 LSTM 模型,CIA-CRF 在 4 个数据集上分别提升了 3.41,2.9,2.25,3.27 个百分点.LSTM 模型将各个词进行 5 分类(标签模式为 BMESO),然而最后的输出可能会出现语法错误的情况,例如标签 E 后的标签为 M,语法错误是 LSTM 模型的性能低于 CIA-CRF 的重要原因.相比于 DTBCSNN+F,CIA-CRF 在 L-14,R-14 数据集上的性能分别提高 3.46 和 0.94 个百分点.DTBCSNN+F 依靠依存句法信息和堆栈神经网络,而本文提出的卷积交互式注意力机制能够更直接捕获到文本中重要的信息(即属性信息),是 DTBCSNN+F 不具备的优势.

在本文所对比的深度学习模型中,还包含了联合学习模型.相比于属性词与情感词的联合抽取模型 MIN 和 GMT,CIA-CRF 在 L-14 和 R-16 数据集上取得了最优的 $F1$ 值,并在 R-15 上取得了与 GMT 可比的性能.MIN 和 GMT 均利用了情感词信息,而本文方法 CIA-CRF 是单一的属性抽取任务,然而在缺少情感词信息辅助的条件下,CIA-CRF 在大部分数据集上仍优于 MIN 和 GMT.

MTCA 为属性词与情感词抽取以及分类的联合学习模型.CIA-CRF 与 MTCA 相比,在 L-14,R-16 数据集上取得更优的效果;而在 R-15 数据集上,CIA-CRF 性能低于 MTCA.经过分析表 3 可知,R-15 的训练集数据量较少.因此,在训练数据偏少时,MTCA 借助情感词抽取以及属性词与情感词分类的辅助信息,从而促进了属性词抽取性能的提升.

本文的模型 CIA-CRF 与基线模型 BiLSTM+CRF 相比,在 4 个数据集上分别提升了 2.21,1.35,2.22,2.21 个百分点.可见,本文提出的类卷积交互式注意力机制应用于属性抽取任务具有一定的优越性.

2.4.2 与全局式注意力模型对比分析

由表 4 可知,在 BiLSTM+CRF 架构下,结合类卷积注意力层并且不引入词的字符级特征(CA-CRF),与基于全局式注意力机制的 GA-CRF 相比,CA-CRF 在 4 个数据集上的性能均得到了提升,分别提升了 0.5,0.83,0.22,0.61 个百分点.经过分析,全局式注意力机制按序(从句首到句尾)动态地对目标词的上下文的所有词分配注意力权重,而距离目标词较远且关联不密切的词就会为目标词的注意力向量带来噪音.为了便于观察评论文本中的注意力分布,我们将一条评论文本样例的每个时刻(t_1 ~ t_{10})注意力得分输出,绘制如图 6 所示的注意力分布图.在图 6 的 t_2 时刻,此时目标词为"service",全局注意力机制为目标词上下文所有的词都分配了注意力权重,而"terrible"这个词距离"service"较远且不相关,却分配了较高注意力权重,从而对目标词"service"的注意力向量带来噪音.

	The	service	is	great	,	but	the	icecream	is	terrible
t_1	0.037	0.132	0.03	0.033	0.028	0.031	0.034	0.543	0.036	0.095
t_2	0.043	0.104	0.035	0.04	0.032	0.036	0.038	0.493	0.044	0.135
t_3	0.036	0.123	0.029	0.032	0.027	0.029	0.032	0.56	0.035	0.098
t_4	0.036	0.119	0.03	0.033	0.028	0.031	0.033	0.56	0.036	0.094
t_5	0.036	0.11	0.029	0.033	0.027	0.03	0.032	0.57	0.036	0.098
t_6	0.036	0.117	0.029	0.032	0.027	0.03	0.032	0.566	0.036	0.096
t_7	0.035	0.123	0.028	0.031	0.027	0.029	0.032	0.566	0.035	0.094
t_8	0.04	0.13	0.033	0.036	0.031	0.034	0.036	0.519	0.039	0.102
t_9	0.035	0.124	0.029	0.032	0.027	0.03	0.032	0.567	0.035	0.09
t_{10}	0.036	0.142	0.03	0.033	0.029	0.031	0.034	0.541	0.036	0.089

Fig. 6 Attention distribution

图 6 注意力分布图

本文提出的类卷积交互式注意力机制中的类卷积注意力层可降低上述噪音,通过设置滑动窗口限制目标词的上下文词的数量,给予窗口内各个词注意力权重,从而获得受噪音干扰较小的注意力向量.

实验结果表明,CA-CRF 性能优于 GA-CRF,在属性抽取上,类卷积注意力层获得的注意力向量更优.

2.4.3 滑动窗口大小设定分析

类卷积注意力层中滑动窗口的大小是重要的超参数,本文将目标词项的上文(下文)词数指定为窗口大小.为了验证滑动窗口大小对实验结果的影响,本文将窗口大小分别设为 2,5,8 进行模型训练,实验过程中保存开发集上 $F1$ 值最优的模型,最后使用最优模型在测试集上进行测试,实验结果如表 5 所示:

Table 5 $F1$ of Different Window Sizes

表 5 不同窗口大小的 $F1$ 对比 %

Model	L-14	R-14	R-15	R-16
CIA-CRF#2	78.61	84.63	70.10	73.99
CIA-CRF#5	79.12	84.91	70.51	73.62
CIA-CRF#8	78.16	84.58	70.32	73.06

从表 5 中可知,当窗口大小为 2 时(CIA-CRF#2),在数据集 R-16 上取得较优的性能;当窗口大小为 5 时(CIA-CRF#5),在 L-14,R-14,R-15 等数据集上性能较优.结合表 3 可发现,R-15 和 R-16 的训练数据平均长度较短,而 L-14 和 R-14 的训练数据平均长度较长.因此,可推测当训练语料的平均长度较短时,应选用较小或稍大的滑动窗口;而当训练语料的平均长度较长时,应选用稍大的滑动窗口.实验中将滑动窗口大小设为 8 时(CIA-CRF#8),在 4 个数据集上的性能均未达到较优的效果,因为较大的滑动窗口会将较多的噪音带入类卷积注意力向量.所以,实验中滑动窗口的大小不能设置过大.由于在大部分数据集上,窗口大小设为 5 都取得了较优的性能.所以,本文在 4 个数据集上统一选择窗口大小为 5 的实验结果作为性能的对比和相应分析.

2.4.4 交互机制对比分析

为了进一步验证类卷积交互式注意力机制中交互注意力层的有效性,本文在 CIA-CRF 的基础上去掉交互注意力层(CIA-CRF-NOI),实验结果与 CIA-CRF 进行对比,如表 6 所示.

从表 6 可发现,在 CIA-CRF 基础上去掉交互注意力层,在 4 个数据集上性能都出现下降,分别下降了 0.94,0.59,0.73,0.6 个百分点.可见,交互注意力层有助于属性词的预测.原因在于,类卷积注意力层按序(从句首到句尾)通过滑动窗口控制每个词(目标词)的上下文词的数量,由于滑动窗口的大小固定,且每个目标词的上下文中与之关联密切的词分布迥异,所以窗口外可能存在与目标词关联密切的词,类卷积注意力向量可进一步优化.在类卷积注意力向量的基础上,从交互注意力层可获得对于类卷积注意力向量而言重要的全局信息,从而有助于属性词的预测.

Table 6 $F1$ of Interactive Attention

表 6 有无交互注意力层的 $F1$ 对比 %

Model	L-14	R-14	R-15	R-16
CIA-CRF-NOI	78.18	84.32	69.78	73.02
CIA-CRF	79.12	84.91	70.51	73.62

2.4.5 字符级特征对比分析

为了验证词的字符级特征对实验结果的影响,本文在 CIA-CRF 的基础上不使用字符级特征(CIA-CRF-NOC),与使用字符级特征的 CIA-CRF 进行对比,对比实验结果如表 7 所示:

Table 7 $F1$ of Character Feature

表 7 有无字符级特征 $F1$ 对比 %

Model	L-14	R-14	R-15	R-16
CIA-CRF-NOC	78.71	84.62	69.62	73.20
CIA-CRF	79.12	84.91	70.51	73.62

从表 7 分析可知,在 CIA-CRF 的基础上去掉字符级特征,在 4 个数据集上性能均下降,分别下降了 0.41,0.29,0.89,0.42 个百分点.对于不加入字符级特征的模型 CIA-CRF-NOC,未登录词的表示采用随机初始化的方法.若未登录词为待抽取的属性词或者与属性词有重要关联的词,随机初始化的方法不利于模型对属性词的预测.与随机初始化的方法相比,从未登录词的本身获得的特征表示更有利于模型对未登录词的识别,进而有利于属性词的预测.表 8 统计了 4 个数据集中登录词和未登录词的数量.

Table 8 Statistics of Login Words and Un-login Words

表 8 登录词和未登录词统计

Dataset	Login Words	Un-login Words
L-14	4 673	337
R-14	5 209	379
R-15	3 522	182
R-16	4 149	249

2.4.6 结合 BERT 的对比分析

预训练模型 BERT[26] 已经在多个自然语言处理任务上取得了优越性能.鉴于此,本节在 4 个数据集上使用 BERT 进行实验.此外,本节还将 BERT 与基线模型 BiLSTM+CRF 结合(BERT+Baseline).同样,本节在 BERT+Baseline 的基础上与类卷积

交互式注意力机制结合(BERT+Baseline+CIA).基于以上,进行实验对比,实验结果如表 9 所示:

Table 9 *F*1 of Combining BERT Models

表 9 结合 BERT 的 *F*1 对比 %

Model	L-14	R-14	R-15	R-16
BERT	78.48	84.78	69.49	74.90
BERT+Baseline	78.66	84.52	67.06	74.13
BERT+Baseline+CIA	79.06	85.53	63.88	72.85

从表 9 可知,在 R-15 和 R-16 数据集上,与 BERT 相比,BERT+Baseline 和 BERT+Baseline+CIA 的性能均下降.结合表 3 分析可知,R-15 和 R-16 的训练数据较少,而 BERT+Baseline 和 BERT+Baseline+CIA 的模型复杂度较高.对于数据量较少的训练数据,复杂度较高的模型容易对其产生过拟合,从而测试性能较差.因此,BERT+Baseline 和 BERT+Baseline+CIA 在 R-15 和 R-16 数据集上,性能均未达到较优.

相比于 R-15 和 R-16,L-14,R-14 的训练语料的数据量较多.在 L-14 和 R-14 数据集上,与 BERT+Baseline 相比,BERT+Baseline+CIA 的性能分别提升 0.4 和 1.01 个百分点.因此,在训练语料的数据量较多的情况下,在 BERT+Baseline 的基础上引入类卷积交互式注意力机制,性能可获得进一步提升,从而也证明了类卷积交互式注意力机制的有效性.

3 总 结

本文提出一种基于类卷积交互式注意力机制的属性抽取方法.该注意力机制包含 2 层注意力,第 1 层是类卷积注意力层,第 2 层是交互注意力层.相比于全局式注意力机制,类卷积注意力层在滑动窗口内为每个词的上下文分配注意力权重,从而获得受噪音干扰较小的类卷积注意力向量.在类卷积注意力层降噪的基础上,通过交互注意力层获得对于类卷积注意力向量而言重要的全局信息.此外,本文提出的模型融入词的字符级特征,字符级特征有助于识别未登录词,从而有助于属性词的预测.实验证明,本文提出的方法在 4 个数据集上性能均有提升.

参 考 文 献

[1] Pontiki M, Galanis D, Pavlopoulos J, et al. SemEval—2014 Task 4: Aspect based sentiment analysis [C] //Proc of the 8th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2014: 27-35

[2] Hu Minqing, Liu Bing. Mining and summarizing customer reviews [C] //Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168-177

[3] Li Zhuang, Feng Jing, Zhu Xiaoyan. Movie review mining and summarization [C] //Proc of the 15th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2006: 43-50

[4] Qiu Guang, Liu Bing, Bu Jiajun, et al. Opinion word expansion and target extraction through double propagation [J]. Computational Linguistics, 2011, 37(1): 9-27

[5] Jakob N, Gurevych I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields [C] //Proc of the 2010 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2010: 1035-1045

[6] Xu Bing, Zhao Tiejun, Wang Shanyu, et al. Extraction of opinion targets based on shallow parsing features [J]. Acta Automatica Sinica, 2011, 37(10): 1241-1247 (in Chinese)
(徐冰, 赵铁军, 王山雨, 等. 基于浅层句法特征的评价对象抽取研究[J]. 自动化学报, 2011, 37(10): 1241-1247)

[7] Liu Pengfei, Joty S, Meng H. Fine-grained opinion mining with recurrent neural networks and word embeddings [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1433-1443

[8] Toh Z, Su Jian. NLANGP at SemEval—2016 task 5: Improving aspect based sentiment analysis using neural network features [C] //Proc of the 10th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2016: 282-288

[9] Wang Wenya, Pan S, Dahlmeier D. Multi-task coupled attentions for category specific aspect and opinion terms co-extraction [J]. arXiv preprint, arXiv: 1702.01776, 2017

[10] Cheng Meng, Hong Yu, Tang Jian, et al. Study on the gated dynamic attention mechanism towards aspect extraction [J]. Pattern Recognition and Artificial Intelligence, 2019, 32(2): 184-192 (in Chinese)
(程梦, 洪宇, 唐建, 等. 面向属性抽取的门控动态注意力机制 [J]. 模式识别与人工智能, 2019, 32(2): 184-192)

[11] Greenberg N, Bansal T, Verga T et al. Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 2824-2829

[12] Jie Zhanming, Lu Wei. Dependency-guided LSTM-CRF for named entity recognition [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 3860-3870

[13] Zhang Han, Guo Yuanbo, Li Tao. Domain named entity recognition combining GAN and BiLSTM-Attention-CRF [J]. Journal of Computer Research and Development, 2019, 56(9): 1851-1858 (in Chinese)

(张晗, 郭渊博, 李涛. 结合 GAN 与 BiLSTM-Attention-CRF 的领域命名实体识别 [J]. 计算机研究与发展, 2019, 56(9): 1851-1858)

[14] Pontiki M, Galanis D, Papageorgiou H, et al. Semeval-2015 task 12: Aspect based sentiment analysis [C] //Proc of the 9th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2015: 486-495

[15] Pontiki M, Galanis D, Papageorgiou H, et al. SemEval-2016 task 5: Aspect based sentiment analysis [C] //Proc of the 10th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2016: 19-30

[16] Yu Jianfei, Jiang Jing, Xia Rui. Global inference for aspect and opinion terms co-extraction based on multi-task neural networks [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(1): 168-177

[17] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C] //Proc of the 2016 Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 260-270

[18] Kim Y. Convolutional neural networks for sentence classification [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1746-1751

[19] Lafferty J, Mcallum A, Pereira F. Conditional random fields probabilistic models for segmenting and labeling sequence data [C] //Proc of the 18th Int Conf on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2001: 282-289

[20] Chernyshevich M. Ihs r&d belarus: Cross-domain extraction of product features using CRF [C] //Proc of the 8th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2014: 309-313

[21] Toh Z, Wang Wenting. Dlirec: Aspect term extraction and term polarity classification system [C] //Proc of the 8th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2014: 235-240

[22] Vicente S, Saralegi X, Agerri R. EliXa: A modular and flexible ABSA platform [C] //Proc of the 9th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2015: 748-752

[23] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation [C] //Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1532-1543

[24] Ye Hai, Yan Zichao, Luo Zhunchen, et al. Dependency-tree based convolutional neural networks for aspect term extraction [C] //Proc of Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2017: 350-362

[25] Li Xin, Lam W. Deep multi-task learning for aspect term extraction with memory interaction [C] //Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 2886-2892

[26] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C] //Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 4171-4186

Wei Zhenkai, born in 1995. Master candidate. His main research interest is information extraction.

Cheng Meng, born in 1994. Master. His main research interest is information extraction.

Zhou Xiabing, born in 1988. PhD. Her main research interest is information extraction.

Li Zhifeng, born in 1994. Master candidate. His main research interest is machine translation.

Zou Bowei, born in 1984. PhD. His main research interest is information extraction.

Hong Yu, born in 1978. PhD. Professor in Soochow University. His main research interest is information extraction.

Yao Jianmin, born in 1971. PhD. Associate professor in Soochow University. His main research interest is machine translation.