计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(6):118−124 [doi: 10.15888/j.cnki.csa.006944]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Multi-head Attention 和 Bi-LSTM 的实体关系分类①

刘　峰[1,2], 高　赛[1,2], 于碧辉[1,2], 郭放达[3]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](东北大学, 沈阳 110819)
通讯作者: 高　赛, E-mail: gsalmh@163.com

摘　要: 关系分类是自然语言处理领域的一项重要任务, 能够为知识图谱的构建、问答系统和信息检索等提供技术支持. 与传统关系分类方法相比较, 基于神经网络和注意力机制的关系分类模型在各种关系分类任务中都获得了更出色的表现. 以往的模型大多采用单层注意力机制, 特征表达相对单一. 因此本文在已有研究基础上, 引入多头注意力机制 (Multi-head attention), 旨在让模型从不同表示空间上获取关于句子更多层面的信息, 提高模型的特征表达能力. 同时在现有的词向量和位置向量作为网络输入的基础上, 进一步引入依存句法特征和相对核心谓词依赖特征, 其中依存句法特征包括当前词的依存关系值和所依赖的父节点位置, 从而使模型进一步获取更多的文本句法信息. 在 SemEval-2010 任务 8 数据集上的实验结果证明, 该方法相较之前的深度学习模型, 性能有进一步提高.

关键词: 关系分类; Bi-LSTM; 句法特征; self-attention; multi-head attention

引用格式: 刘峰,高赛,于碧辉,郭放达.基于 Multi-head Attention 和 Bi-LSTM 的实体关系分类.计算机系统应用,2019,28(6):118–124. http://www.c-s-a.org.cn/1003-3254/6944.html

Relation Classification Based on Multi-Head Attention and Bidirectional Long Short-Term Memory Networks

LIU Feng[1,2], GAO Sai[1,2], YU Bi-Hui[1,2], GUO Fang-Da[3]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Northeastern University, Shenyang 110819, China)

Abstract: Relation classification is an important subtask in the field of Natural Language Processing (NLP), which provides technical support for the construction of knowledge map, question answer systems, and information retrieval. Compared with traditional relational classification methods, deep learning model-based methods with attention have achieved better performance in various relation classification tasks. Most of previous models use one-layer attention, which cause single representation of the feature. Therefore, on the basis of the existing works, the study introduces a multi-head attention, which aims to enable the model to obtain more information about sentence from different representation subspaces and improve the model's feature expression ability. Otherwise, based on the existing word embedding and position embedding as network input, we introduce dependency parsing feature and relative core predicate dependency feature to the model. The dependency parsing features include the dependency value and the location of the dependent parent node position for the current word. The experimental results on the SemEval-2010 relation classification task show that the proposed method outperforms most of the existing methods.

① 收稿时间: 2018-12-13; 修改时间: 2019-01-08; 采用时间: 2019-01-15; csa 在线出版时间: 2019-05-25

Key words: relation classification; Bi-LSTM; syntax feature; self-attention; multi-head attention

信息抽取是自然语言处理领域的一项重要任务, 其目标是从普通的非结构化文本中抽取易于机器或程序理解的结构化信息, 从而将互联网上大量的信息存储成一个庞大的知识库, 提供给用户查看或者为其它自然语言处理任务提供服务. 随着信息技术的高速发展, 网络上的信息越来越庞大, 信息抽取任务就变得愈发重要.

关系抽取作为信息抽取的一个重要组成部分, 它旨在从语义层面发现实体之间的关系. 关系抽取可以使用一组三元组来描述<Entity1,Relation,Entity2>, 其中 Entity1 和 Entity2 表示实体, Relation 表示两个实体之间的关系. 例如“<e1>叶莉</e1>是<e2>姚明</e2>的妻子”. 其中“<e1>”和“</e1>”这两个符号声明第一个实体为“叶莉”, “<e2>”和“</e2>”则声明第二个实体为“姚明”. 可以看出, 两个实体之间的关系是"配偶". 在无监督或半监督学习领域, 关系抽取是指从没有任何预先确定的实体和关系类别中提取事实以及关系短语; 在监督学习领域, 关系抽取又可以看作一项关系分类任务, 是指将包含已知实体对的文本的实体关系分类到一组已知的关系类别上. 本文的研究是在具有既定关系和已知实体对的数据集上进行关系抽取任务, 因此本文的关系抽取任务就是一项关系分类任务.

传统的关系分类方法常用的有两种, 基于规则的方法和基于特征向量的方法. 基于规则的方法需要领域专家的介入且需要人工构建大量的匹配规则, 可扩展性差. 基于特征的方法需要人工构建大量的特征, 费时费力, 且人工提取的特征都停留在词法和句法层面, 模型无法很好地捕获文本的语义特征. 近年来, 随着深度学习的发展, 神经网络模型开始应用在各类关系分类任务上,并取得出色表现. 本文在此研究基础上, 提出基于多头注意力机制 (multi-head attention) 和双向长短时记忆网络 (Bi-LSTM) 相结合的实体关系分类模型. 本文主要贡献如下:

(1) 引入 aulti-head Attention. 它是 self-attention 的一种拓展, 能够从多个不同维度计算 attention, 从而使模型在不同子空间学习特征.

(2) 模型的输入在已有的词向量和位置向量的基础上, 进一步引入依存句法特征和相对核心谓词依赖特征作为输入, 可以使模型更好地捕获句法信息, 进一步提高模型分类的精度.

1 相关研究

目前, 已有的关系分类方法包括: 基于规则的方法、基于特征向量的方法、基于核函数的方法和基于深度学习模型的方法.

基于规则的方法需要依赖领域专家, 通过构建大量的模式匹配规则进行关系分类, 适合于特定领域的关系分类任务. Aone[1]等通过人工构建匹配规则开发了 REES 系统, 该系统可识别 100 多种关系. Humphreys[2]等对文本进行句法分析, 通过构建复杂的句法规则来识别实体间的关系. 基于规则的方法需要领域专家的指导, 耗时耗力, 且系统可移植性差.

基于特征向量的方法需要人工构造特征, 然后将特征转化为向量, 利用机器学习算法构建模型, 将特征向量作为模型的输入对实体对之间的关系进行分类. Kambhatla[3]等人通过结合词汇特征、句法特征和语义特征, 利用最大熵模型作为分类器, 在 ACE RDC2003 的评测数据集上, 最终分类的 F 值达到了 52.8%. 车万翔[4]等人通过引入实体类型、两个实体的出现顺序、实体周围的 w 个词等特征, 利用支持向量机 (SVM) 作为分类器, 在 ACE RDC2004 的评测数据集上, 最终分类的 F 值达到了 73.27%. 基于机器学习的方法依赖于人工构造特征, 其效果的好坏也严重依赖于特征选取的好坏, 且为达到较高的分类性能往往需要从不同层次构造大量特征.

基于核函数的方法不需要显示构造特征, 因此避免了人工构造特征的过程. 它将文本的字符串或者文本的句法树作为输入实例, 使用核函数计算实例间的相似度来训练分类器. 在关系分类任务中使用核函数的方法最早是由 Zelenko[5]等人引入的,他们在文本的浅层解析表示上定义核函数, 并将核函数与支持向量机 (SVM) 和投票感知器学习算法相结合. 实验表明, 该方法取得了良好的效果.

近年来, 随着深度学习的兴起, 越来越多的研究工作都尝试使用神经网络模型去解决问题, 从而避免显式的人工构造特征的过程. Liu CY 等人[6]在关系分类

任务中最早尝试使用卷积神经网络自动学习特征. 它建立了一个端到端 (End-to-End) 的网络, 该网络利用同义词向量和词法特征对句子进行编码, 实验结果表明, 该模型在 ACE 2005 数据集上的性能比当时最先进的基于核函数的模型的 F 值高出 9 个百分点. Zeng DJ 等人[7]也使用了卷积神经网络模型来进行关系分类, 他们使用了预先在大型未标记语料库上训练的词向量 (Word Embedding), 并首次将位置向量 (Position Embedding) 引入模型的输入. 最终该模型在 SemEval-2010 任务 8 的评测数据集上的 F 值达到了 82.7%.

卷积神经网络 (CNN) 虽然在关系抽取任务中取得了不错的表现, 然而 CNN 不适合具有长距离依赖信息的学习. 循环神经网络 (RNN) 适用于解决具有长距离依赖的问题, 但是它存在梯度消失问题, 对上下文的处理就受到限制. 为了解决这个问题, Hochreiter 和 Schmidhuber 在 1997 年提出长短时记忆网络 (LSTM), 该网络通过引入门控单元来有效缓解 RNN 的梯度消失问题. 另外, 近年来基于神经网络和注意力机制 (attention) 相结合的模型也被广泛应用在关系分类任务上. 注意力机制是对人类大脑注意力机制的一种模拟, 最早应用在图像处理领域, Bahdanau 等人[8]最早将其应用在机器翻译任务上. 此后注意力机制就被广泛地应用到各种 NLP 任务中. Zhou P 等人[9]提出一种用于关系分类的神经网络 ATT-BLSTM. 该模型利用长短时记忆网络对句子进行建模, 并结合自注意力机制 (self-attention) 来进一步捕捉句子中重要的语义信息. 通过计算 self-attention, 可以得到句子内部词之间依赖关系, 捕获句子内部结构. 本文的研究在文献[9]工作的基础上, 引入多头注意力机制 (multi-head attention), 其本质是进行多次 self-attention 计算, 可以进一步提高实体关系分类精度.

2 基于 Multi-head Attention 和 Bi-LSTM 的关系分类算法

本文采用双向长短时记忆网络 (Bi-LSTM) 对文本特征进行建模. 在将词向量和相对位置向量作为网络层输入的基础上, 进一步考虑将依存句法特征和相对核心谓词依赖特征引入网络输入层. 将这两个特征引入输入层的原因是:

(1) 依存句法分析可以很好地揭示文本句法结构, 并且反映出两个实体之间直接或间接的关系特征.

(2) 大量研究表明, 对一个句子的所有谓词, 核心谓词对于识别实体边界、承接实体关系起着至关重要的作用[10]. 因此每个词与核心谓词的相对依赖也是一种隐含特征, 这种依赖关系必然也能反映出实体间的关系特征.

同时在网络输出层引入 multi-head attention. Multi-head attention 由 Vaswani[11]等人提出, 基于 Self-Attention. Self-Attention 通过计算每个词和所有词的注意力概率来捕获句子的长距离依赖. 所谓 multi-head, 就是进行多次 Self-attention 计算, 每次计算时使用的映射矩阵不同, 最后将每一次计算结果进行拼接, 作为最终 multi head 计算结果. 容易看出 multi head attention 和单头 self-attention 相比, 它可以学习多个映射器, 进而从不同维度, 不同子空间来表征特征. 最后通过将多个特征进行拼接进行特征融合, 可以使模型进一步提高特征表达能力. 文献[11]中的实验结果表明, 使用单头注意力机制可以学习得到句子内部词的某些长距离依赖关系, 而 multi-head attention 除了能够加强这种学习能力以外, 甚至能够理解句子的句法和语义结构信息. 因此本文引入 multi-head attention 思想, 来进一步提高模型建模能力, 从而提高实体关系分类的精度.

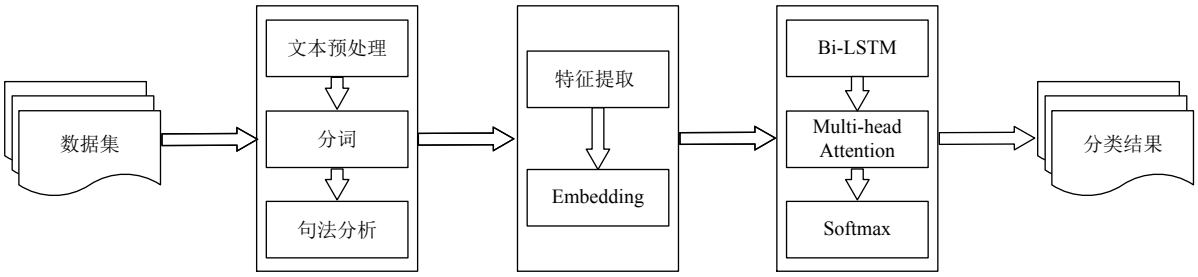

图 1 模型框架图

本文的模型包含以下 5 个部分, 模型结构图如图 2 所示.

(1) 文本预处理、特征提取.

(2) Embedding 层: 将网络输入的各种特征全部映射为低维向量表示.

(3) Bi-LSTM 层: 使用 Bi-LSTM 对输入信息进行建模, 获取高层特征表示.

(4) Multi-head attention 层: 进行多次 self-attention 计算, 并将多次计算结果进行拼接和线性映射, 获取最终句子级特征表示.

(5) 输出层: 采用 SoftMax 函数作为分类器, 将上一步得到的特征向量作为输入, 可以得到最终的关系类别.

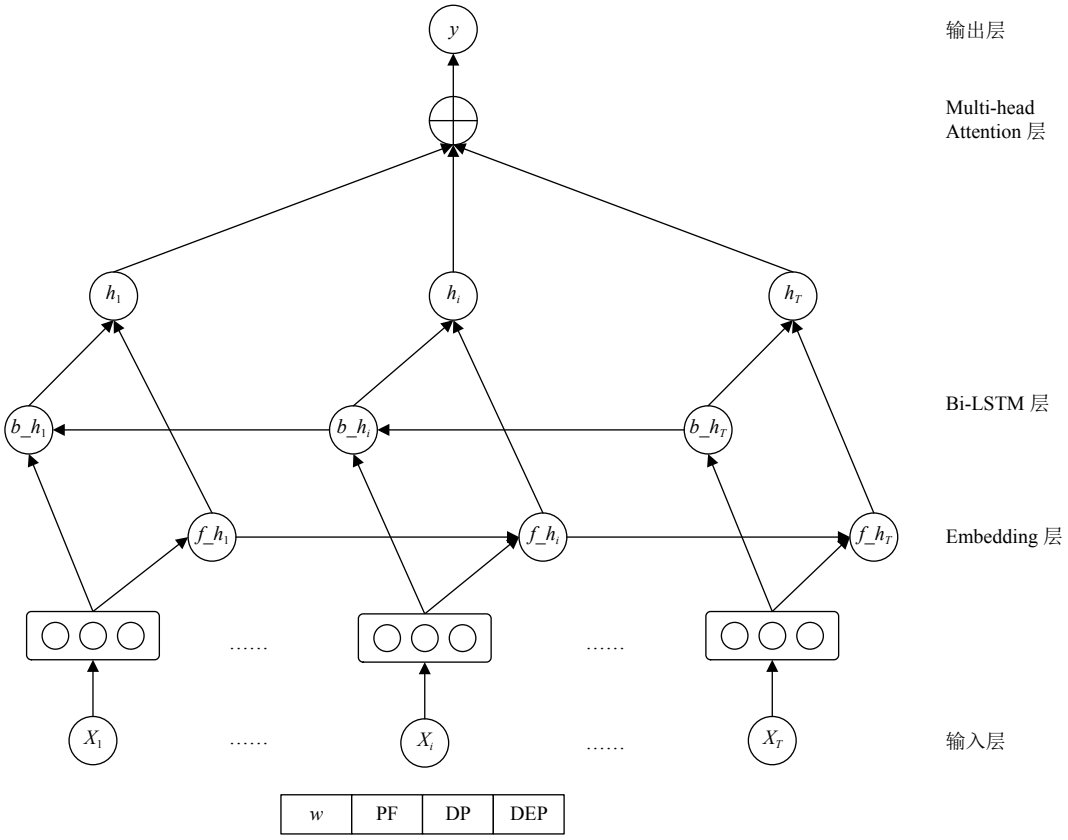

图 2 模型结构图

2.1 文本预处理、特征提取

以"<e1>叶莉</e1>是<e2>姚明</e2>的妻子"为例, 使用哈工大的 LTP 对句子进行分词和依存句法分析, 结果如下图所示, 抽取以下三个特征:

(1) 相对位置特征 PF. 即句子中每个词分别到实体 1 和实体 2 的距离. 如例句中两个实体分别是“叶莉”、“姚明”. 每个词到实体 1“叶莉”的距离 PF1={0, 1, 2, 3, 4}; 每个词到实体 2“姚明”的距离 PF2={−2, −1, 0, 1, 2}

(2) 依存句法特征 DP. 特征 DP 包含两部分 DP_NAME 和 DP_PAR. DP_NAME 要获取每一个词在句子中的依存句法属性值, 那么例句的 DP_NAME={SBV, HED, ATT, RAD, VOB}; DP_PAR 要获取每一个词所依赖的词在句子中的索引值, 那么例句的 DP_PAR={2, 0, 5, 3, 2}

(3) 相对核心谓词依赖特征 DEP. 根据句子中每个词与核心谓词是否存在依赖关系, 将 DEP 特征取值分为三类: DEP_S(核心谓词本身), DEP_C(核心谓词子节点), DEP_O(其它). 容易看出例句的核心谓词为“是”, 那么例句的 DEP={DEP_C,DEP_S,DEP_O, DEP_O, DEP_C}.

2.2 Embedding 层

假定句子 S 由 T 个词组成, $S=\{w_1, w_2, \cdots, w_T\}$, 对于每个词$w_i$都要提取五种特征, 用$e_i^j$表示, 其中$1 \leqslant j \leqslant 5$.

每个特征所对应的特征向量矩阵分别为: $\{W^{word}, W^{pf}, W^{dp\_name}, W^{dp\_par}, W^{dep}\}$. $W^{word} \in R^{d^w \times |V|}$, $W^f \in R^{d^v \times |V_f|}$, d^w 是词向量的维度, $|V|$ 表示数据集词汇量大小. $f \in \{pf, dp\_name, dp\_par, dep\}$, d^v 是相应特征向量的维度, V_f 表示特征 f 取值类别个数. W^{word} 使用一个预训练好的词向量矩阵[12], 其余特征向量矩阵都采用随机初始化的方式赋予初始值. 使用式 (1) 对每个词的各个特征进行 Embedding, 得到每个特征的向量化表示.

$$e_i^j = W^{kj} v_i^j \tag{1}$$

其中, v_i^j 表示词 w_i 在第 j 个特征处取值的对应索引值. 将每个特征向量 e_i^j 进行拼接得到每个词的最终向量化表示 e_i. 最终, 句子 S 在 Embedding 层的输出为:

$$embs = \{e_1, e_2, \cdots, e_T\} \tag{2}$$

2.3 Bi-LSTM 层

LSTM 是 RNN 的一种变体, 它通过引入门控单元克服 RNN 长期依赖问题从而缓解梯度消失. 一个 LSTM 单元由三个门组成, 分别是输入门 i_t, 遗忘门 f_t 和输出门 o_t. 以特征 $embs = \{e_1, e_2, \cdots, e_T\}$ 作为输入, 将 t 作为当前时刻, h_{t-1} 表示前一时刻隐层状态值, c_{t-1} 表示前一时刻细胞单元状态值, 计算第 t 时刻词对应的 LSTM 各个状态值:

$$i_t = \sigma(W_{xi} e_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf} e_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \tag{4}$$

$$g_t = \tanh(W_{xc} e_t + W_{hc} h_{t-1} + W_{cc} c_{t-1} + b_c) \tag{5}$$

$$c_t = i_t g_t + f_t c_{t-1} \tag{6}$$

$$o_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_t + b_o) \tag{7}$$

$$h_t = o_t \tanh(c_t) \tag{8}$$

通过以上计算, 最终得到 t 时刻 LSTM 隐层状态的输出值 h_t. 在本文中使用的是 Bi-LSTM. 将前向 LSTM 中 t 时刻隐层状态值记为 f_h_t, 将后向 LSTM 中 t 时刻隐层状态的输出值记为 b_h_t, 则最终 Bi-LSTM 第 t 时刻输出值为:

$$h_t = f\_h_t + b\_h_t \tag{9}$$

2.4 Multi-head Attention 层

Multi-head attention 本质就是进行多次 self-attention 计算, 它可以使模型从不同表征子空间获取更多层面的特征, 从而使模型能够捕获句子更多的上下文信息. Multi-head attention 模型结构如图 3 所示.

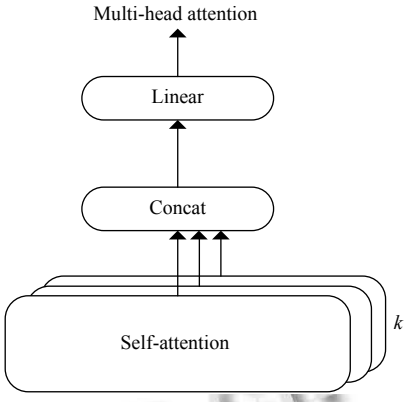

图 3 Multi-head attention

(1) 单次 self-attention 计算. 使用符号 H 表示一个矩阵, 它由 Bi-LSTM 层所有时刻输出向量组成 $[h_1, h_2, \cdots, h_T]$. 使用符号 r 表示该层最终的输出值, 计算过程如下:

$$\partial = SoftMax(w^T \tanh(H)) \tag{10}$$

$$r = H\partial^T \tag{11}$$

其中, $H \in R^{d^h \times T}$, d^h 是隐藏层节点数, w 是一个参数向量. w, ∂ 和 r 的维度分别是 d^h, T, d^h. 经过 self-attention 计算, 可以得到单次 attention 输出特征值为:

$$h^* = \tanh(r) \tag{12}$$

(2) Multi-head attention 计算. 即进行 k 次 self-attention 计算. 在计算过程中, 针对式 (10), 在每次使用 H 时, 需要先将 H 进行一次线性变换[11], 即 $W_i^h H$, 其中 $W_i^h \in R^{d^h/k \times d^h}$, $i \in \{1, 2, \cdots, k\}$. 这样, 每次在进行单次 self-attention 计算时, 都会对 H 的维度进行压缩, 且 multi-head attention 计算可以并行执行. 另外, 本文使用的是乘法注意力机制, 乘法注意力机制在实现上可以使用高度优化的矩阵乘法, 那么整体计算成本和单次注意力机制的计算成本并不会相差很大, 同时又提升了模型的特征表达能力. 使用式 (10)~(12) 进行 k 次计算, 注意每次计算使用的 w 均不相同. 将结果 h^* 进行拼接和线性映射, 得到最终结果 h_s:

$$h_s = w_s \otimes Concat(h_1^*, h_2^*, \cdots, h_k^*) \tag{13}$$

其中, 向量 w_s 的维度是 $k \times d^h$, $\otimes$ 表示逐元素点乘.

2.6 输出层

在本文中, 关系分类为一个多分类问题. 使用 SoftMax 函数计算每一个类别的条件概率, 然后选取条件概率最大值所对应的类别作为预测输出类别. 计算

过程如下:

$$p(y'|S) = SoftMax(W_o h_s + b_o) \tag{14}$$

$$y = \arg\max_{y'} p(y'|S) \tag{15}$$

其中, $W_o \in R^{c\times kd_w}$, c 表示数据集的类别个数. 目标函数是带有 L2 正则化的类别标签 y 的负对数似然函数:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i') + \lambda\|\theta\|_2^2 \tag{16}$$

其中, m 是样本的个数, $t_i \in R^m$是一个关于类别的 one-hot 向量, y_i'是 SoftMax 的输出概率向量, λ是 L2 正则化因子

3 实验结果与分析

3.1 实验数据

本次实验采用 SemEval-2010 任务 8 的数据集. 该数据集共包含 10 种关系类别, 其中有 9 种是明确的关系类别, 一种是未知类别“Other”. 数据集中共有 10 717 条人工标注实体和关系类别的数据, 包括 8000 条训练数据, 2717 条测试数据. 关系类别如表 1 所示.

表 1 关系类别

关系类型标号	关系类型名称
1	因果关系
2	整体-部分关系
3	实体-目标关系
4	实体-来源关系
5	生产者-产品关系
6	会员-组织关系
7	实体-主体关系
8	内容-包含关系
9	工具-使用者关系
10	其它关系

3.2 实验评价指标

在本次实验中采用官方评测标准 $F1$ 值 ($F1$-Score) 作为模型性能评价指标. 表 2 为分类结果的混淆矩阵.

表 2 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

在计算$F1$值之前, 需要先计算查准率 P、查全率 R, 计算公式如下:

$$P = \frac{TP}{TP+FP} \tag{17}$$

$$R = \frac{TP}{TP+FN} \tag{18}$$

根据 P、R 值计算最终的$F1$值:

$$F1 = \frac{2PR}{P+R} \tag{19}$$

3.3 参数设置

本文神经网络模型的优化方法采用 Adam, 其学习率设置为 1.0, 激活函数使用 relu 函数, 隐层节点数 d^h设置为 300, 每个单词嵌入向量大小d^w为 50, 输入的 batch 大小为 50. 另外, 为了缓解过拟合现象, 在目标函数中加入 L2 正则化项, 正则化因子取值为10^{-5}, 同时引入 dropout 策略. 将 dropout 应用在 Embedding 层、Bi-LSTM 层, 经过多轮实验 (采用 5 折交叉验证), 当丢码率 (dropout rate) 分别为 0.3, 0.3, 模型可以达到一个比较好的性能. Multi-head 层中的参数 k 的值过大或过小都不好, 参考文献[11]的实验, 取[1, 2, 4, 6, 10, 15, 30]作为候选值 (k 要能被d^h整除), 采用 5 折交叉验证方法评估模型性能, 实验结果如下表所示. 易知, 最终当 k 值为 4 的时候模型可以达到一个较好的性能. 单次 self-attention 要比 $k=4$ 时 multi-head attention 的效果差, 但随着 k 值的不断增加, 模型性能会下降. 故最终选取 k 的值为 4.

表 3 k 值实验结果

k	$F1$(%)
1	83.9
2	84.3
4	85.4
6	85.0
10	84.3
15	83.8
30	83.1

3.4 实验结果

为将本文模型与其它模型效果进行对比实验, 所有模型均采用同一数据集, 关系类别个数为 10. RNN 模型、ATT-LSTM 模型的输入词向量和位置向量、网络隐层节点数、网络激活函数、模型优化方法等均与本文实验中的参数设置保持一致. 另外 CNN 中与本文无关的参数设置参考其原论文. 实验结果如表 4.

CNN: 该模型是文献[7]提出的. 使用 CNN 对句子进行建模, 同时引入位置特征和词汇特征, 使用 SoftMax 作为分类器. 最终实验结果$F1$值达到 80.3%.

RNN: 该模型是文献[13]提出的. 使用双向 RNN 来进行关系分类, 使用 SoftMax 作为分类器. 最终实验结果$F1$值达到 81.5%.

ATT-LSTM: 该模型文献[9]提出. 使用双向 LSTM 对句子进行建模, 并引入自注意力机制, 使用 SoftMax 作为分类器. 最终实验结果 $F1$ 值达到 83.4%.

表 4 实验结果

模型	$F1$(%)
CNN	80.3
RNN	81.5
ATT-LSTM	83.4
本文方法	85.4

以上四种模型相比, 本文提出的方法最终$F1$值达到 85.4%, 均高于以上三种模型. 本文模型与以上三种模型相比, 在 embedding 层, 进一步引入了句法层面的信息. 与 CNN 和 RNN 方法相比, 本文神经网络结构采用双向 LSTM. 双向 LSTM 相比 CNN 更能捕获具有长期依赖的信息, 更适合处理文本序列; 与 RNN 相比, LSTM 通过引入门控机制, 缓解了模型的梯度消失问题. 与 ATT-LSTM 模型相比, 本文的模型将单层 self-attention 改为 multi-head attention. 综上所述, 本文方法在 embedding 层融入了更加丰富的句法特征, 通过使用双向 LSTM 使模型学到更多具有长期依赖的上下文信息, 在最后的 attention 层, 通过使用 multi-head attention 进一步提高了模型的特征表达能力. 通过实验验证, 本文方法进一步提高了实体关系分类模型的精度.

4 结语

本文从现有的基于深度学习模型的关系抽取方法出发, 使用 Bi-LSTM 和 multi-head attention 机制对文本进行建模, 同时为了使模型更好地学习到文本句法结构信息, 进一步引入句法结构特征和相对核心谓词依赖特征. 在公共评测语料上的实验结果证明该方法相较于其他深度学习模型性能有进一步提升. 未来的工作可考虑如何进一步改进 attention 以及如何将模型应用到无监督关系抽取研究上.

参考文献

1 Aone C, Ramos-Santacruz M. REES: A large-scale relation and event extraction system. Proceedings of the 6th Conference on Applied Natural Language Processing. Seattle, WA, USA. 2000. 76–83.

2 Humphreys K, Gaizauskas R, Azzam S, *et al*. University of Sheffield: Description of the LaSIE-Ⅱ system as used for MUC-7. Proceedings of the 7th Message Understanding Conference. VA, USA. 1998.

3 Kambhatla N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. Proceedings of ACL 2004 on Interactive Poster and Demonstration Sessions. Barcelona, Spain. 2004. 22. 71–85.

4 车万翔, 刘挺, 李生. 实体关系自动抽取. 中文信息学报, 2005, 19(2): 1–6. [doi: 10.3969/j.issn.1003-0077.2005.02.001]

5 Zelenko D, Aone C, Richardella A. Kernel methods for relation extraction. The Journal of Machine Learning Research, 2003, 3: 1083–1106.

6 Liu CY, Sun WB, Chao WH, *et al*. Convolution neural network for relation extraction. Proceedings of the 9th International Conference on Advanced Data Mining and Applications. Hangzhou, China. 2013. 231–242.

7 Zeng DJ, Liu K, Lai SW, *et al*. Relation classification via convolutional deep neural network. Proceedings of the 25th International Conference on Computational Linguistics. Dublin, UK. 2014. 2335–2344.

8 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2014.

9 Zhou P, Shi W, Tian J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016. 207–212.

10 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取. 中文信息学报, 2014, 28(6): 183–189. [doi: 10.3969/j.issn.1003-0077.2014.06.026]

11 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA, USA. 2017. 5998–6008.

12 Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. 2010. 384–394.

13 Zhang DX, Wang D. Relation classification via recurrent neural network. arXiv: 1508.01006, 2015.