控制与决策
Control and Decision

结合注意力机制的循环神经网络复述识别模型

李旭， 姚春龙， 范丰龙， 于晓强

引用本文:
李旭, 姚春龙, 范丰龙, 等. 结合注意力机制的循环神经网络复述识别模型[J]. 控制与决策, 2021, 36(1): 152−158.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0638

结合注意力机制的循环神经网络复述识别模型

李 旭†, 姚春龙, 范丰龙, 于晓强

(大连工业大学 信息科学与工程学院，辽宁 大连 116034)

摘 要: 传统基于深度学习的复述识别模型通常以关注文本表示为核心,忽略了对多粒度交互特征的挖掘与匹配. 为此,建模文本交互空间,分别利用双向长短时记忆网络对两个候选复述句按条件编码,基于迭代隐状态的输出,通过逐词软对齐的方式从词、短语、句子等多个粒度层次推理并获取句子对的语义表示,最后综合不同视角的语义表达利用softmax实现二元分类. 为解决复述标注训练语料不足,在超过580 000句子对的数据集上利用语言建模任务对模型参数无监督预训练,再使用预训练好的参数在标准数据集上有监督微调. 与先前最佳的神经网络模型相比,所提出模型在标准数据集MSRP上准确率提高2.96 %, F_1 值改善2 %. 所提出模型综合文本全局和局部匹配信息,多粒度、多视角地描述文本交互匹配模式,能够降低对人工特征工程的需求,具有良好的实用性.

关键词: 自然语言处理；复述识别；循环神经网络；双向长短时记忆；注意力机制；无监督预训练

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0638 **开放科学(资源服务)标识码(OSID):**

引用格式: 李旭,姚春龙,范丰龙,等. 结合注意力机制的循环神经网络复述识别模型[J]. 控制与决策, 2021, 36(1): 152-158.

Recurrent neural networks based paraphrase identification model combined with attention mechanism

LI Xu†, YAO Chun-long, FAN Feng-long, YU Xiao-qiang

(School of Information Science and Engineering，Dalian Polytechnic University，Dalian 116034，China)

Abstract: The traditional paraphrase identification models based on deep learning usually focus on text representation and ignore the mining and matching of multi-granular interaction features. To address the problem, we propose a recurrent neural network model with word-by-word attention mechanism. In this paper, the word embeddings are inputted into the recurrent neural networks, and the two candidate paraphrase sentences are conditionally encoded via two bidirectional. Based on the output of the iterative hidden states, the sentence-pair representation is obtained from global matching and fine-grained reason via soft-alignment of words and words in the two sentences. Finally, for classification, we use a softmax layer over the output of a non-linear projection of the output vector into the target space of the two classes. The labeled training set for paraphrase identification is small in comparison with the high complexity of the task. In order to make full use of the training data, we use a language modeling task to unsupervised pre-train the neural network parameters on the corpora of more than 580,000 pairs of sentences. This is followed by a fine-tuning stage, where we adapt the model to a specific task with labeled data. Compared with the previous state-of-art neural network model, the accuracy and the F_1 score of our model are improved by 2.96 percent and 2 percent on the MSRP data set respectively. The proposed model combines multiple semantic expressions of text from different perspectives and describes the multi-granular matching pattern. It is an end-to-end differentiable system that reduces manual feature engineering efforts, and has good practicability.

Keywords: natural language processing；paraphrase identification；recurrent neural networks；bidirectional long short-term memory；attention mechanism；unsupervised pre-training

0 引 言

复述识别主要判断两个措辞不同的句子是否表达相同或者几乎相同的语义,是自然语言理解中的一个核心问题. 实用高效的复述识别模型会对包括机器翻译、自动文摘、文本查重、自动问答、文本分类、信息检索等在内许多自然语言问题的研究和应

收稿日期: 2019-05-10；**修回日期:** 2019-07-22.

基金项目: 国家重点研发计划专项项目(2017YFC0821003-3)；辽宁省高等学校基本科研项目(2017J049)；辽宁省自然科学基金项目(20180550395)；辽宁省教育厅青年科技人才“育苗”项目(J2020113).

†通讯作者. E-mail: lixu102@aliyun.com.

用产生较大帮助.

传统的复述识别模型需要大量的工程技术和专业领域知识人工设计特征,将原始数据转换成适当的内部特征表示或特征向量,通过学习系统对输入样本进行二元分类. 然而,人工定义和抽取特征不仅费时费力,而且这些特征总是针对特定的任务而设计,在一个任务上表现很好的特征很难应用到其他自然语言处理任务上,很大程度上限制了模型的泛化能力. 究其原因在于复述识别面临的挑战主要来源于自然语言的歧义性、多元性和复杂性. 近几年,随着深度学习取得的不断进展和计算机硬件计算能力的迅速提升,复述识别问题的研究渐渐从传统的人工设计分类特征向深度学习匹配模型转移.

在深层神经网络中,各层特征都不是人工设计,而是通过一种学习过程从数据中学到. 深度学习是一种特征学习方法,能够将原始数据通过一些简单的但是非线性的模型转换成更高层次更抽象的表示. 对于分类任务,高层次更抽象的表示能够强化输入数据的区分能力,同时削弱不相关因素. 然而,现存的基于深度学习的复述识别模型通常具有以下不足:

1) 传统模型侧重于文本的语义表示,如ARC-I[1]和Qiu等[2]提出的CNTN模型. 首先利用深度学习方法生成每个句子的向量表示,然后计算两个句子语义向量之间的匹配程度. 虽然该方法语义表达简单,计算速度快,但是它将两个句子之间的交互延迟到各自的表示矩阵,每个句子的语义表示是在彼此独立的情况下生成的,忽略了对交互特征的挖掘与匹配. 此外,编码器在压缩句子过程中存在信息损失,使得传统模型对两个句子的语义向量进行相似度计算时,丧失了较多的语义细节和交互计算.

2) 传统模型确定两个句子语义等价时通常只关注一个粒度级别上的匹配,如Tien等[3]提出的MaxLSTM-CNN模型仅考虑句子层次上的语义匹配. Madnani等[4]使用机器翻译自动评价中常用的NIST、BLEU等作为分类特征,但是这些特征仅仅是低层次的局部特征,文本全局表示的高层次特征未被考虑. 由于文本是以层次化的方式组织起来的,在句义匹配时需考虑词、短语、句子等多个粒度层次上的语义推理和匹配.

3) 虽然存在大量充足的未标注自然语言文本,但是针对复述识别任务的标注数据是稀缺的,传统模型仅使用小规模数据训练具有大量参数的网络模型,容易过度拟合训练数据,导致模型泛化性能下降.

为解决上述不足,本文提出一种结合注意力机制的循环神经网络复述识别模型. 主要贡献在于: 1)所建模型基于多粒度语义推理的文本交互匹配模式,在交互信息的基础上进一步挖掘匹配特征. 不仅保留了每个句子抽象的个体发展空间,而且能实现对两个句子之间匹配模式的提取和融合. 所提出的逐词注意力机制综合文本局部细粒度和全局匹配信息,能够多粒度地描述文本交互匹配模式,为输出提供有效决策. 2) 在大规模语料上无监督预训练模型参数不仅可以解决复述标注数据不足的瓶颈问题,防止过拟合,增强监督模型的泛化能力,而且可以使模型充分学习并有效挖掘出隐含在大量数据背后含义不明显的分类特征,提高模型的鲁棒性. 此外,无监督训练得到的词向量相当于在构建数据集上对词嵌入进行微调,更加适用于复述识别任务. 3) 标准数据集上的实验结果验证了所提出模型的有效性,该模型是一个端对端的分类系统,能降低对人工特征工程的需求,具有良好的实用性.

1　复述识别模型

用数学符号对复述识别问题形式化如下:

输入: $(s_1, s_2), s_1 \in S$和$s_2 \in S$为两个自然句,S为自然句集合;

输出: $y \in Y, Y = \{0, 1\}$,其中0表示不为复述,1表示互为复述;

训练样本: $D = \{(s_1^{(1)}, s_2^{(1)}, y^{(1)}), \ldots, (s_1^{(i)}, s_2^{(i)}, y^{(i)}), \ldots, (s_1^{(n)}, s_2^{(n)}, y^{(n)})\}, i = \{1, 2, \ldots, n\}$,其中$s_1^{(i)} \in S$, $s_2^{(i)} \in S, y^{(i)} \in Y$.

复述识别模型的目标是在训练样本中自动学习匹配函数$f : S \times S \to Y$,使得对于测试数据上的任意输入$(s_1', s_2')(s_1' \in S$且$s_2' \in S)$能够预测出描述两个句子是否表达相同语义的类别标签$y'(y' \in Y)$.

所提出的复述识别模型架构如图1所示,分为编码器、基于语言模型的无监督预训练组件和面向复述识别任务的有监督分类器3部分.

1.1　编 码 器

编码器包括输入层、隐藏层和注意力层,目标是基于交互空间推理和学习句子对的语义表示.

1.1.1　输 入 层

大量研究已经证明,使用从大规模未标注语料库中学习的词嵌入可以有效提高自然语言处理任务的性能. 然而,不同的词嵌入模型捕获的语言属性不同,基于词袋上下文模型倾向于反映域知识,而基于复述关系的模型则善于捕捉词语之间的语义相似性. 因此,对于不同的自然语言处理任务,应该学习并使用面向特定任务的词嵌入. 基于此, 本文首先使用预训

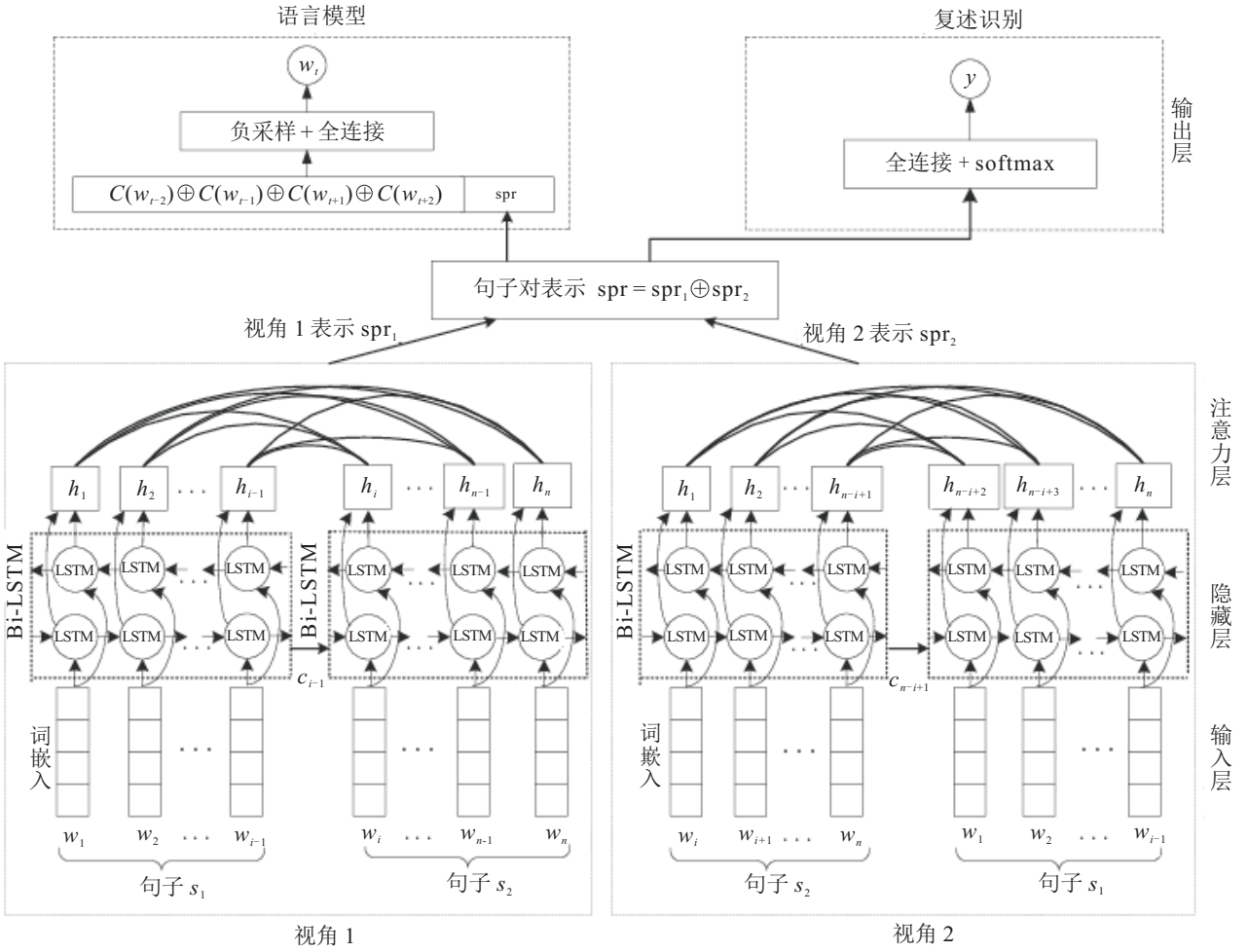


图 1 模型架构

练好的GloVe (global vectors) 词向量[5]初始化网络模型,然后针对语言建模任务将词向量作为待确定的模型参数,通过微调训练得到.

1.1.2 隐 藏 层

在隐藏层,模型使用Bi-LSTMs对输入序列编码,使其转化成一个固定长度的向量. 如图1所示,Bi-LSTM包括输入、前向层、后向层和输出,令x_t、h_t分别表示时间步t上的输入向量和隐状态输出,前向层和后向层分别为LSTM,共同连接输出. 前向层从第1时刻到最后时刻正向计算,获得并保存每个时刻前向隐藏层的输出,后向层沿着最后时刻到时刻1反向计算并保存每个时刻后向隐藏层的输出. LSTM每个记忆细胞中被设置控制信息流入和流出3种类型的门:输入门i、遗忘门f和输出门o. 通过门机制使信息选择性地通过,从而实现应该记忆的信息一直传递,不应该记忆的信息被门截断. 将Bi-LSTM输出中前向$\overrightarrow{h_t}$和后向$\overleftarrow{h_t}$对应时刻的输出结果相加作为该时刻最终的隐状态输出h_t.

本文模型关注语义交互匹配模式,因此模型同时读入两个句子,而不是将每个句子独立地映射到一个语义空间中. 如图1所示,在隐藏层两个候选复述句分别由两个Bi-LSTM按条件编码,即第1个Bi-LSTM的最终隐藏状态作为第2个Bi-LSTM的初始状态. 为了防止过拟合,定义LSTM后在细胞外部包裹上dropout,使LSTM记忆细胞中各个门对数据流控制更加优秀,增加学习效果.

1.1.3 注意力层

本文提出的逐词注意力机制如图1所示. 令Y为由第1个Bi-LSTM输出向量$[h_1 \quad h_2 \quad \dots \quad h_i \quad h_L]$ $(i \in [1, L])$组成的矩阵,$Y \in \boldsymbol{R}^{k \times L}$. 其中:$k$为记忆细胞的隐藏神经元个数,$L$为文本最大长度,$h_i$为第$i$时刻的隐状态输出. 当第2个Bi-LSTM处理第2个句子时,在每一时刻分别计算当前输出与第1个Bi-LSTM输出向量中哪些元素具有语义相关性,并按照输出对之间的语义相关程度打分,分值越高表示语义相关程度越高,对应上下文受到的关注程度也应该越高. 在任意时刻$t, t \in (L+1, N)$,使用如下公式计算注意力得分:

$$S_t = \tanh(W^y Y + W^h h_t \otimes e_L). \tag{1}$$

其中:不同的W(用不同上角标区分)表示不同的权

重矩阵, W^y、$W^h \in \boldsymbol{R}^{k\times k}$; e_L 是长度为 L 的全1向量; $h_t \otimes e_L$ 表示将列向量 h_t 重复 L 次, 形成一个 L 列的矩阵. t 时刻注意力权重 $\alpha_t(\alpha_t \in \boldsymbol{R}^L)$ 计算如下:

$$\alpha_t = \mathrm{softmax}(W^{\mathrm{T}} S_t). \tag{2}$$

使用Bi-LSTM获取各个时间步的隐状态输出后, 将带有注意力权重的隐状态信息汇总, 结合上一时刻的句子对加权表示, 生成当前时刻的句子对加权语义表示. 在任意时刻 t, 句子对加权语义向量 $r_t(r_t \in \boldsymbol{R}^k)$ 计算如下:

$$r_t = Y\alpha_t^{\mathrm{T}} + \tanh(W^r r_{t-1}), \tag{3}$$

其中 r_t 不仅依赖于当前时刻的注意力表示, 而且依赖于 r_{t-1} 告知模型在前一时刻哪些文本片段被关注以及被关注的程度. 时序序列中每一时刻的语义匹配信息被保存并传递给下一时刻, 最后时刻的加权语义向量 r_N 包含了整个序列的全局匹配信息. 因此, 该注意力机制能够基于局部匹配构建全局匹配, 增强整体语义匹配的质量.

1.1.4　双视角结构

Bi-LSTM从两个方向同时读取序列能够解决单向LSTM处理数据时存在位置偏见的问题, 改善了编码性能. 受其启发, 本文提出一种双视角结构, 使用相同的网络结构和权重分别从不同视角扫描句子对, 同时获得多个语义表示. 如图1所示, 在视角1中, 句子对的输入顺序为 s_1 在前 s_2 在后, 在视角2中两个句子顺序互换, s_2 在前 s_1 在后. 在计算每个视角文本语义表示时, 综合考虑以下两方面因素: 一是最后时刻综合全局上下文信息的隐状态输出, 二是通过逐词交互计算方式从局部细粒度推理中获得的加权语义向量. 计算公式如下:

$$\mathrm{spr}_j = \tanh(W^x h_{N_j} + W^r r_{N_j}). \tag{4}$$

其中: W^r、$W^x \in \boldsymbol{R}^{k\times k}$; $j \in [1,2]$; $\mathrm{spr}_j \in \boldsymbol{R}^k$; h_{N_j} 和 r_{N_j} 分别表示第 j 个视角下最后时刻的输出向量和加权语义表示. 将两个视角的语义表示向量 spr_1 和 spr_2 相加形成句子对的最终表示spr.

1.2　无监督预训练

为了建模纷繁复杂的复述现象, 获得良好性能, 本文提出一种针对复述识别的半监督训练方法, 以无监督形式学习句子层次的文本表示, 有效减少对监督学习的依赖. 本文将语言模型拼接到复述识别模型的最高层表示上, 首先以语言模型作为目标任务训练得到最优的模型参数组合, 然后使用预训练得到的参数对复述识别任务初始化, 在复述标注数据上微调分类器. 训练方法流程如图2所示.

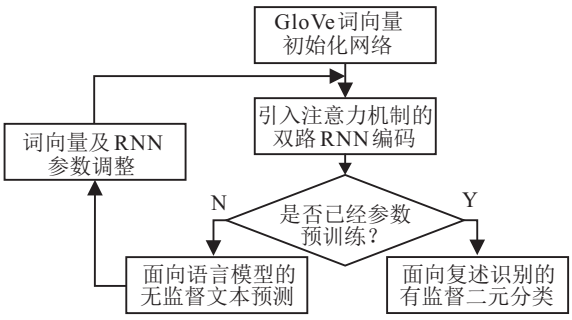


图2　训练方法流程

统计语言模型是基于语料库计算一个词语序列概率的模型. 基于神经网络的语言模型在实际应用中常根据上下文词语预测目标词, 最大化如下对数似然函数:

$$L = \sum_{w\in C} \log P(w|\mathrm{Context}(w);\Theta). \tag{5}$$

其中: C 为语料; $\mathrm{Context}(w)$ 为词 w 的统一上下文, 即 w 周边的词的集合; Θ 为网络待确定参数集.

由于无需任何标注信息即可实现训练, 语言模型可以充分利用大规模自然语言文本进行建模, 学习丰富的语义知识. 本文使用类似于CBOW (continuous bag-of-words)[6] 的模型结构, 模型包括输入层和输出层. 本文模型与CBOW模型的不同之处在于, CBOW的输入层仅为当前词 w_t 的局部上下文 w_{t-2}, $w_{t-1}, w_{t+1}, w_{t+2}$, 本文语言模型的输入层除了包括当前词的上述局部语境外, 还包括综合全局上下文信息的句子对表示spr. 将当前词周边的前后各 n 个词的词向量作累加求和(取 $n=2$, 即 $C(w_{t-2})\oplus C(w_{t-1})\oplus C(w_{t+1})\oplus C(w_{t+2})$), 再与句子对表示spr拼接起来作为一个全连接层的输入去生成目标词的预测表示, 如图1所示. 为了优化训练, 使用负采样技术, 即每次随机采样一些不相关的词放在样本中作为负样本, 训练时只要实现目标词(正样本)比其他噪声(负样本)概率大即可.

设 $\{(x_i, y_i^k)\}_{i=1}^n$ 为一个样本数据, y_i^0 表示目标词, $y_i^1, \ldots, y_i^k$ 表示噪声词, 则成本函数为

$$J = -\frac{1}{n}\sum_{i=1}^{n}\left[\log P(y_i^0|x_i,\Theta) + \sum_{t=1}^{k}\log P(y_i^t|x_i,\Theta)\right]. \tag{6}$$

通过最小化损失训练模型, 词向量以及RNN (recurrent neural networks)模型参数捆绑在一起, 无监督训练完成后两者同时得到.

1.3　有监督分类器

将句子对表示spr连接一个全连接层, 再利用softmax函数生成归一化概率值, 最大概率的索引即

为二元分类的类别标签y,有

$$y = \text{argmax}(\text{softmax}(W^s \text{spr})). \tag{7}$$

其中: $W^s \in \boldsymbol{R}^{k\times 2}, \text{spr} \in \boldsymbol{R}^k$.

为避免过拟合,复述识别模型在交叉熵损失函数中加入L_2正则化,成本函数为

$$J = -\frac{1}{n}\sum_{i=1}^{n}[y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})] + \frac{\lambda}{2n}\|\theta\|_2^2. \tag{8}$$

其中: $y^{(i)}$和$\hat{y}^{(i)}$分别为第i个样本的真值和预测值; θ为权重矩阵中的所有列向量; λ为超参数,用来控制正则化的强度.

2 实验结果与分析

2.1 数据集与评价指标

MSRP (microsoft research paraphrase)[7]是复述问题的一个公开标准数据集. 数据集共包含5 801条句子对,将数据集切分为训练集、验证集和测试集3部分. 训练集包含3 481条句子对,其中复述关系成立的正例2 320对,复述关系不成立的负例1 161对. 验证集和测试集分别包含1 160条句子对,其中正例790对,负例370对.

实验采用准确率和F_1分值作为客观评价指标. 准确率为分类正确的样本数占总样本数的比例; F_1分值为综合衡量精确率和召回率的评价指标,计算精确率和召回率的调和平均值.

2.2 实验设置与对比结果

本文预训练数据来源于未标注的MSRP、SICK (sentences involving compositional knowledge)[8]以及用于识别文本蕴含任务的SNLI (stanford natural language inference)[9]语料,共包含585 880条英文句子对. 设置句子的最大长度为70, 利用pad_sequences( )方法对变长序列自动尾部填充. 词嵌入维度$d = 300$. 设置滑动窗口half_window_size $= 2$, 即局部语境为目标词前后各两个词. 每次采样的负样本数量为20. 为获得更好的收敛速度和收敛性,模型使用Adam优化算法以减小成本函数为优化目标. 在每次模型优化中,分别调整LSTM神经元数目、L_2正则化强度、dropout概率以及批大小的值,在验证集上找到最优参数组合,然后在测试集中评估模型性能. 具体训练时,应根据每次配置信息的运行结果调整相应的参数大小,不断迭代更新,逐步优化模型参数.

设置LSTM神经元数目$k = 200$、dropout $= 0.2$、batchsize $= 40, \lambda = 0.03$,本文模型在MSRP测试集上准确率为81.56 %, F_1值为86.72 %. 本文模型与其他系统在MSRP数据集上的实验结果如表1所示.

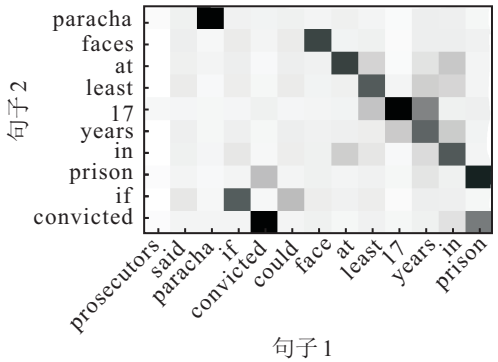
表1 MSRP数据集上实验结果 %

模型	准确率	F_1值
baseline (ARC-I(2014))[1]	69.6	80.3
ARC-II (2014)[1]	69.9	80.9
Bi-CNN-MI- (2015)[10]	72.5	81.4
Bi-CNN-MI (2015)[10]	78.1	84.4
MV-LSTM (2016)[11]	75.4	82.8
Match-SRNN (2016)[11]	74.5	81.7
Subword+LM (2018)[12]	-	84.0
M-MaxLSTM-CNN (2018)[3]	78.1	84.5
OpenAI GPT (2018)[13]	-	82.3
GLUE (2018)[14]	75.1	83.7
He et al (2015)[15]	78.6	84.7
本文模型	**81.56**	**86.72**

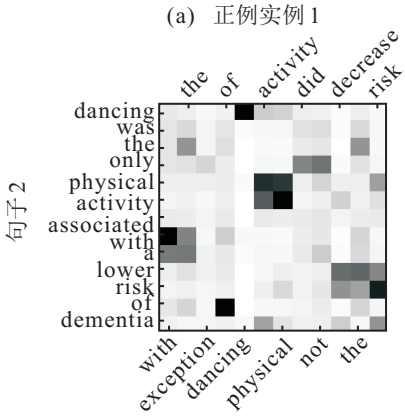
本文使用Hu等[1]提出的ARC-I作为基准线. ARC-I与ARC-II[1]均不需要先验语言知识,利用CNN (convolutional neural network)自动学习特征,适用于多种自然语言处理任务,然而其性能还有待提高. Bi-CNN-MI-和Bi-CNN-MI[10]使用CNN提取及匹配多粒度特征,其中Bi-CNN-MI在大规模数据集上进行了参数预训练,相比于未使用预训练的Bi-CNN-MI-模型,准确率显著提高. 庞亮等[11]提出了多视角循环神经网络MV-LSTM和基于二维循环神经网络的Match-SRNN,上述两个模型在MSRP数据集上的运行结果如表1所示. Lan等[12]进行句子对建模时使用预训练的词嵌入和子词嵌入组合,在MSRP数据集上F_1值达到84 %,准确率未报道. M-MaxLSTM-CNN[3]使用LSTM对输入的多方面级词嵌入编码生成句子嵌入,再通过多级比较学习句子间的语义关系. 该模型与本文模型均无需人工构建特征,但前者准确率和F_1值均明显低于本文模型. GPT[13] (generative pre-training)和多任务基准测试及分析平台GLUE (general language understanding evaluation)[14]虽然使用了语言模型预训练的方法,然而两个模型过于通用,对多粒度文本交互匹配模式考虑不足,在复述识别任务上的表现不如本文模型. He等[15]提出的多视角句子相似度计算模型在先前模型中性能最好,模型输入除了使用公开的GloVe、Paragram词嵌入外,还额外使用200维的词性嵌入. 本文模型在未使用任何人工定义及提取特征的情况下,与多视角句子相似度计算模型相比准确率提高2.96 %, F_1值改善2 %,在所有深度模型中表现最佳.

2.3 注意力可视化

下面讨论和分析逐词注意力机制对文本中不同上下文信息表示的关注程度. 在训练集的正例和负例样本中随机选择两对文本实例,注意力可视化分别如图3和图4所示. 图中每个小方格表示对应的输出向量学习到的注意力权重,其颜色表示注意力权重的大小,权重越大,小方格对应的颜色越深.



(a) 正例实例1

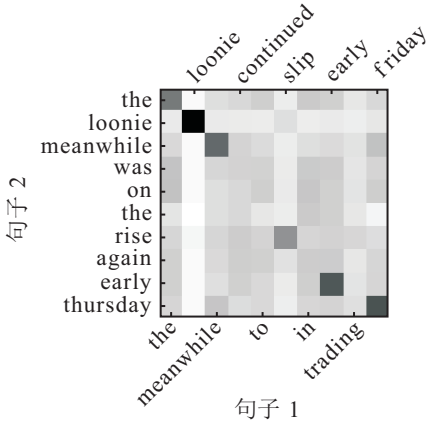


(b) 正例实例2

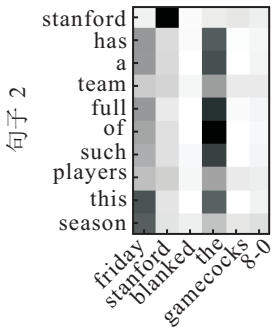
图 3　正例注意力可视化

由图3可见,图3(a)中,注意力主要集中在相同的词语对和短语对上,句子2中每个单词均出现在句子1中,句子2仅是句子1单词的重新排序,该实例表明注意力模型很容易检测出词序变换形式的复述. 图3(b)中,句子对中共有的单词或短语权值较大,此外,文本对“with a”“with the”“lower risk”“decrease the risk”也获得了较多的注意力关注,这表明模型能有效识别文本中的同义词短语,而“dementia”(阿尔茨海默病)“physical”“only”“did not”获得的注意力关注表明该模型在某种程度上能挖掘出文本间潜在的更深层次的语义联系.

由图4可见,图4(a)中,注意力权重的分布相对均衡,“thursday”与“friday”,“rise”与“slip”获得了相对较多的注意力,表明该模型能专注于文本间语义表达的矛盾之处,为分类器输出提供有效的支持和决策. 图4(b)中,除了共有单词stanford获得较多关注



(a) 负例实例1



(b) 负例实例2

图 4　负例注意力可视化

外,注意力基本都集中在定冠词“the”上,究其原因在于句子1与句子2表达的语义完全无关,表明此时句子对表示更多地取决于循环神经网络最后时刻的输出向量,而不是加权的注意力表示.

2.4 模型简化测试

为了验证所提出模型结构的有效性,进行模型简化测试. 对于模型的主要组件(按条件编码、注意力机制、多视角结构和无监督预训练)每次去除一个,在数据集上重新训练和测试. 表2为简化测试结果与最终结果(见表1)之间在准确率和F_1值上的差异.

表2　模型简化测试结果　　%

ID	简化组件	准确率	F_1值
1	去除Bi-LSTM间的按条件编码	−1.37	−1.01
2	去除注意力机制层	−3.15	−2.36
3	去除多视角结构	−1.46	−1.13
4	去除无监督预训练,输入词嵌入使用300维Glove词向量	−3.34	−2.3

由表2可见,去除两个Bi-LSTM之间按条件编码,准确率和F_1值分别降低了1.37%和1.01%,表明按条件编码能够使信息交互流动起来,在处理句子2时,不需要编码其全部语义,只需要识别出与句子1中语义相同或矛盾的单词或短语即可. 进行注意

力的简化测试时,去除注意力层,仅使用按条件编码后最后时刻的隐状态输出作为句子对表示. 与使用注意力机制相比,模型准确率和F_1值分别降低了3.15 %和2.36 %,上述实验结果验证了所提出的注意力机制的有效性. 注意力层通过对上下文信息的加权筛选,使网络细粒度推理出不同文本之间的相互关系,能有效提高分类器的性能. 多视角结构的简化测试结果与最终实验结果之间的差异,表明多视角扫描能有效弥补编码器在压缩句子过程中的信息损失,为分类器带来更丰富的细节信息和语义表示,从而使模型达到更好的效果. 相比于去除以语言模型为附属任务的无监督预训练,复述识别模型的准确率和F_1值分别降低3.34 %和2.3 %,两者之间的显著差异充分表明了预训练语言模型的有效性. 使用预训练参数初始化模型不仅可以加快收敛速度,而且在构建数据集上训练得到的词语分布式表征更关注文本之间的语义相似性和可区分特征,有效提升了复述识别模型的整体性能.

3 结 论

复述识别在自然语言处理任务中有广泛的应用,为此本文提出了一种结合注意力机制的循环神经网络复述识别模型,将无监督预训练语言模型和有监督微调分类任务相结合,能够多粒度、多视角地描述文本交互匹配模式,有效避免和减少人工定义和抽取特征工程,具有良好的实用性. 在今后的研究中,计划将该模型应用于情感分析、自动问答等其他自然语言处理任务.

参考文献(References)

[1] Hu B T, Lu Z D, Li H, et al. Convolutional neural network architectures of matching natural language sentences[J]. 2015, arXiv.1503.03244.

[2] Qiu X P, Huang X J. Convolutional neural tensor network architecture for community-based question answering[C]. Proc of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires: AAAI, 2015: 1305-1311.

[3] Tien N H, Le N M, Tomohiro Y. Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity[J]. 2018, arXiv: 1805.07882.

[4] Madnani N, Tetreault J. Re-examining machine translation metrics for paraphrase identification[C]. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montreal: ACL, 2012: 182-190.

[5] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014: 1532-1543.

[6] Mikolov T, Sutskever I, Chen K. Distributed representations of words and phrases and their compositionality[J]. 2013, arXiv: 1310.4546.

[7] Quirk C, Brockett C, Dolan W B. Monolingual machine translation for paraphrase generation[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Barcelona Spain: ACL, 2004: 142-149.

[8] Marelli M, Menini S, Baroni M. A SICK cure for the evaluation of compositional distributional semantic Models[C]. Proceedings of the Conference on Language Resources and Evaluation. Reykjavik, 2014: 216-223.

[9] Bowman S R, Angeli G, Potts C. A large annotated corpus for learning natural language inference[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015: 632-642.

[10] Yin W P, Schutze H. Convolutional neural network for paraphrase identification[C]. Proceedings of the North American Chapter of the Association for Computational Linguistics. Denver: ACL, 2015: 901-911.

[11] 庞亮, 兰艳艳, 徐君. 深度文本匹配综述[J]. 计算机学报, 2017, 40(4): 985-1003.
(Pang L, Lan Y Y, Xu J. A survey on deep text matching[J]. Chinese Journal of Computers, 2017, 40(4): 985-1003.)

[12] Lan W W, Xu W. Character-based neural networks for sentence pair modeling[C]. Proceedings of the North American Chapter of the Association for Computational Linguistics. New Orleans: ACL, 2018: 157-163.

[13] Radford A, Narasimhan K, Salimans T. Improving language understanding by Generative Pre-training[EB/OL]. [2018-06-11]. https://blog.openai.com/language-unsupervised.

[14] Wang A, Singh A, Michael J. GLUE: A multi-task benchmark and analysis platform for natural language understanding[J]. 2018, arXiv: 1804. 07461.

[15] He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015: 1576-1586.

作者简介

李旭(1980−), 女, 副教授, 博士, 从事深度学习、文本挖掘等研究, E-mail: lixu102@aliyun.com;

姚春龙(1971−), 男, 教授, 博士, 从事数据挖掘、智能计算等研究, E-mail: yaocl@dlpu.edu.cn;

范丰龙(1972−), 男, 副教授, 从事信息系统、数据挖掘等研究, E-mail: fanfl@dlpu.edu.cn;

于晓强(1974−), 男, 教授, 从事智能计算、大数据分析与挖掘等研究, E-mail: tigeryxq@dlpu.edu.cn.

(责任编辑: 郑晓蕾)