DOI:10.11918/201910183

层次化神经网络模型下的释义识别方法

袁 蕾,高 曙,郭 淼,袁自勇

(武汉理工大学 计算机科学与技术学院,武汉 430000)

摘 要: 释义识别技术(Paraphrase Identification, PI)被广泛用于问答系统、抄袭检测、个性化推荐等领域.针对已有释义识别方法缺乏有效的特征提取机制问题,提出了一种新的释义识别模型.与传统"编码-匹配"模式不同,采用"编码-匹配-提取"模式,通过添加特征提取层进一步提取分类信息.所提出模型由6层组成:输入层、嵌入层、编码层、匹配层、特征提取层、输出层.在编码层,采用基于注意力机制的上下文双向长短期记忆网络对文本上下文进行编码,充分利用句子的前向和逆向两个方向的上下文信息;在匹配层,通过多种矩阵运算,从不同角度获得句子对匹配信息;在特征提取层,利用Xception网络以便更有效地从匹配结果中提取分类信息.此外,本文采用多特征融合的方法,将GloVe预训练的词向量、字符向量和附加特征向量的连接作为最终的词向量,较普通的词向量携带更丰富的语义信息.实验结果表明,所构建的模型在Quora和SemEval-2015 PIT两个公开数据集上(分别作为大型数据集和中小型数据集的代表)都达到了竞争性效果.

关键词: 自然语言处理;释义识别;Xception;注意力机制;双向长短期记忆网络

中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2020)10-0175-08

Paraphrase identification based on hierarchical neural network

YUAN Lei, GAO Shu, GUO Miao, YUAN Ziyong

(School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430000, China)

Abstract: Paraphrase identification is widely used in question answering system, plagiarism detection, and personalized recommendation. Since the existing paraphrase identification techniques are lack of effective feature extraction mechanism, a new paraphrase model was proposed. Different from previous works which normally adopt the "encoding-matching" mode, the proposed model adopts the "encoding-matching-extraction" mode by adding feature extraction layer to better acquire classification information. The proposed model is consisted of six layers: input layer, embedding layer, encoding layer, matching layer, feature extraction layer, and output layer. The encoding layer utilizes contextual bi-directional long short-term memory network (BiLSTM) with self-attention to encode context of sentences, which can make full use of contextual information in both forward and reverse directions of a sentence. The matching layer uses several matrix operations to get sentence pair matching information from different angles. The extraction layer chooses Xception as the feature extractor to better extract classification information from the matching results. Moreover, this paper combines GloVe word vectors, character vectors, and additional feature vectors as the final embeddings, which carries richer information than ordinary pretrained embeddings. Results show that the proposed model achieved competitive results on two public datasets: Quora Question Pairs (as a representative of large datasets) and SemEval-2015 PIT (as a representative of small and medium datasets).

Keywords: natural language processing; paraphrase identification; Xception; attention mechanism; BiLSTM

随着互联网的发展和个人移动终端的普及,互联网上产生的信息以爆炸方式增长.这些信息数量庞大、种类繁杂,且大部分是以短文本(或句子)的方式存在的,包括Twitter、微博的推文,电商网站的用户评价等.这些短文本数据具有稀疏性、实时性、不规范性等特点,导致人工处理这些海量的短文本信息极其困难.对用户生成的噪声文本进行释义识别是自然语言处理、信息检索、文本挖掘领域的重要任务,对查询排名、剽窃检测、问答、文档摘要等领域也起到了重要作用[1].最近,由于需要处理语言变异的问题,释义识别任务已经在自然语言处理领域中获得了极大的关注.

释义识别,又称复述检测,通常被形式化为二进制分类任务:对于给定的两个句子,确定它们是否具有相同的含义,具有相同含义的句子称为释义对,而具有不同含义的句子称为非释义对[2].

传统的释义识别方法主要关注文本的特征,包括字面特征、语法特征、语义特征等.但这些方法存

收稿日期: 2019-10-27
基金项目: 国家自然科学基金(51679180)
作者简介: 袁 蕾(1997—),女,硕士研究生
通信作者: 高 曙,gshu@whut.edu.cn

在准确率不高和受到语料库限制导致适应性差等缺点. 随着神经网络发展,专家学者们陆续提出了各种基于神经网络的释义识别模型. 这些基于神经网络的释义识别模型大大提高了识别的准确率,但仍存在一些问题:易受到数据集限制,在大型数据集上表现良好的模型,常常在小型数据集上表现较差等. 同时,现有神经网络释义识别模型大多采用“编码-匹配”模式,对句子对进行编码、匹配操作以后,结果被直接用于分类,没有充分利用匹配结果中的信息. 针对这些问题,本文提出了一种面向释义识别的层次化神经网络模型,它采用了“编码-匹配-提取”模式,编码层使用基于注意力的上下文双向长短期记忆力网络(Attention Based Contextual Bi-directional Long Short-Term Memory Network, ABC-BiLSTM)作为编码器,获取前向和逆向两个长短期记忆力网络(Long Short-Term Memory Network, LSTM)所有隐藏层状态,并且通过注意力机制(Attention Mechanism)提取权重信息;匹配层利用多种矩阵运算获得匹配结果;特征提取层则利用 Xception 作为提取器,以便进一步从句子匹配结果中提取分类特征.

1 国内外研究现状

近年来,国内外相关学者在释义识别领域投入了大量的研究. 识别两个句子是否是释义对,即是识别二者是否足够相似,包括字面上的相似和语义上的相似. 现有的释义识别方法主要有基于特征的方法和基于神经网络的方法.

基于特征的方法主要关注文本的特征,包括 n-gram 重叠特征[3]、语法特征[4]、语言特征[5-6]、基于维基百科的语义网络[7]、知识图[8]等. 该类方法通过提取文本对的特征,然后通过计算特征向量的相似度,判断两个文本是否是释义对. 计算特征向量的相似度方法有余弦相似度、欧式距离以及词移距离等方法.

基于神经网络的方法有两种,一种是通过神经网络计算词向量,然后计算词向量的距离得到文本相似度,判断是否是释义对. 如黄江平等使用神经网络训练词向量,并使用改进的 EMD 方法计算向量间的语义距离,获得文本释义关系[9]. 另一种是通过神经网络模型直接输出文本是否是释义对,本质上是一种分类算法. 常用的神经网络模型有卷积神经网络(Convolutional Neural Network,CNN)、递归神经网络(Recurrent Neural Network, RNN)、注意力机制等. 在这些模型的基础上,学者们提出了各种适用于释义识别的神经网络模型. 包括 Wang 等提出的 BiMPM,通过双向长短期记忆网络(Bi-directional Long Short-Term Memory Network, BiLSTM)编码句子,在两个方向上匹配来自多个角度的编码结果[10];Chen 等的 ESIM 模型,使用两层 BiLSTM 和自注意力机制,将编码后结果通过平均池化层和最大池化层,输入决策层分类[11];Kim 等设计的一种具有密集连接的互注意力循环神经网络 DRCN,主要由单词表示层,注意力机制连接的 RNN 编码层和交互预测层组成[12]等.

综上,基于特征和基于神经网络的释义识别方法,或者受到语料库限制,或者缺乏特征提取机制,或者模型准确率对数据集大小较敏感,有待进一步提升。因此,本文设计了面向释义识别的层次化神经网络模型,通过增加特征提取层并在相关层提取更丰富的语义和分类信息,从而克服以上问题.

2 问题定义

对于给定长度为 p 的句子 $\boldsymbol{A}=(a_1,\cdots,a_p)$ 和长度为 q 的句子 $\boldsymbol{B}=(b_1,\cdots,b_q)$,求分类结果 $y\in\{0,1\}$. $y=0$ 表示两个句子含义不同(是释义对),$y=1$ 表示两个句子含义相同(非释义对).

3 面向释义识别的层次化神经网络

本文提出的面向释义识别的层次化神经网络(Hierarchical Paraphrase Identification Network, HPIN)模型是一种分层结构,由输入层、嵌入层、编码层、匹配层、特征提取层、输出层组成. 图 1 显示了该模型的整体结构. 与已有的释义识别神经网络模型不同,HPIN 采用“编码-匹配-提取”模式,在“编码-匹配”模式基础上,添加了特征提取层,以便从匹配结果中提取更多分类信息. HPIN 各层的概述如下.

1)输入层用于将句子转换为向量形式,即用不同的数字表示不同的单词. 该层的输入是句子对,输出是向量对.

2)嵌入层使用密集分布向量表示输入句子的每个单词,向量之间的距离表示语义的相似程度. 该层对预训练词向量(包含可训练和不可训练两种)、字符向量和附加特征向量进行连接,并作为最终词向量. 嵌入层的输入是向量对,输出是词向量矩阵对.

3)编码层用于学习句子的上下文信息. 编码层采用基于注意力机制的上下文双向长短期记忆力网络,能够获取前向和逆向两个 LSTM 中所有单元的隐藏状态. 该层的输入是词向量矩阵对,输出是编码矩阵对.

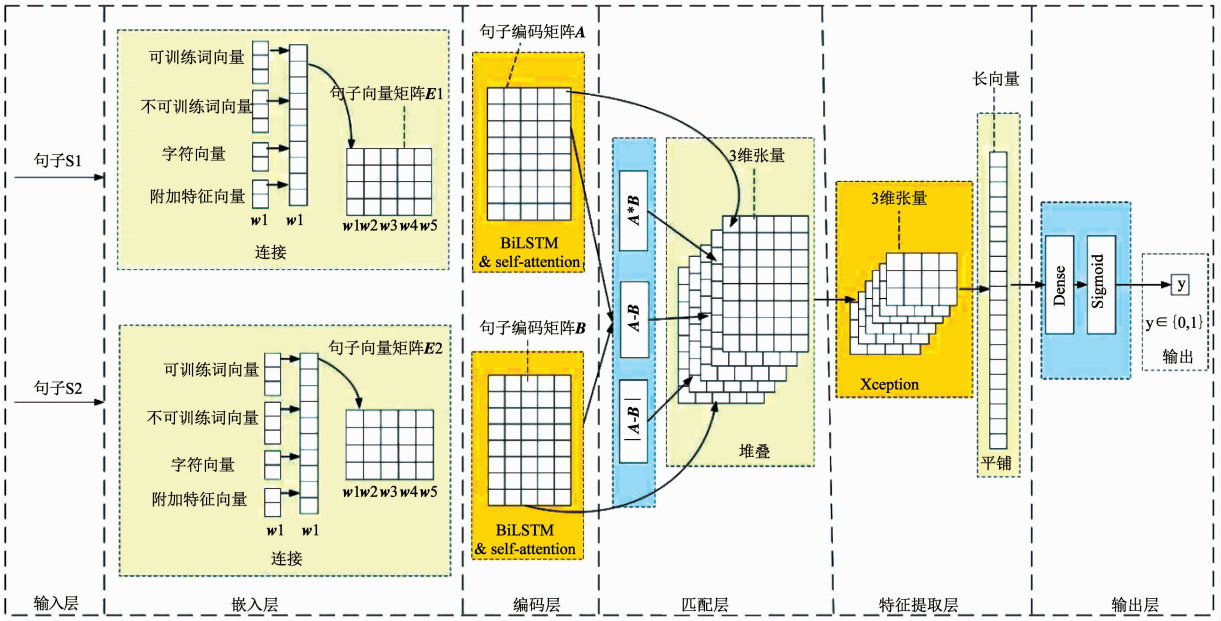

图 1　面向释义识别的层次化神经网络模型

Fig. 1　Hierarchical neural network model for paraphrase identification

4)匹配层对编码结果进行多种矩阵运算,包括矩阵减法、矩阵相减再按位求绝对值、矩阵按位乘法,并且与编码矩阵对堆叠,生成三维张量. 该层的输入是由编码层生成的编码矩阵对,输出是三维张量.

5)特征提取层用于提取匹配层输出的三维张量的语义特征. HPIN 使用 Xception 作为编码器以便更有效地从匹配结果中提取分类信息. 该层的输入是由匹配层生成的三维张量,并将 Xception 的输出平铺成一个长向量,作为特征提取层的输出.

6)输出层由密集层和 sigmoid 函数组成,用于判断句子对是否为释义对. 该层的输入是特征提取层生成的长向量,输出是二进制值,1 代表是释义对,0 代表非释义对.

3.1　多特征融合的词向量表示

在嵌入层,每个单词被表示为一个密集分布的向量,整个句子因而被表示为词向量矩阵. 使用可训练词向量、不可训练词向量、字符向量和附加特征向量的串联作为最终的词向量.

1)可训练词向量和不可训练词向量. 使用 840B 通用语料预训练的 GloVe 作为词向量. 可训练词向量指在训练过程中会被更新的词向量,不可训练词向量指在训练过程中不会被更新的词向量. 在嵌入层,两种词向量都会被使用.

2)字符向量. 使用一维卷积核过滤字符向量. 单词的字符卷积特征在时间维度上最大池化获得向量. 字符特征能够为一些词汇表外(Out-of-Vocabulary, OOV)的单词提供额外信息.

3)附加特征向量. 通过“附加特征筛选实验及分析”选取合适的附加特征组合,从而得到附加特征向量. 嵌入层使用的附加特征有 Wordnet 相似度和词性标注.

最终的词向量由可训练词向量、不可训练词向量、字符向量、附加特征向量连接而成,具体可表示为

$$\boldsymbol{E}(\boldsymbol{P}) = [\boldsymbol{t}(\boldsymbol{P}), \boldsymbol{u}(\boldsymbol{P}), \boldsymbol{c}(\boldsymbol{P}), \boldsymbol{f}(\boldsymbol{P})]. \quad (1)$$

式中:P 为句子,$E(P)$ 为句子 P 的词向量矩阵,$\boldsymbol{t}(\boldsymbol{P})$为可训练词向量,$\boldsymbol{u}(\boldsymbol{P})$为不可训练词向量,$\boldsymbol{c}(\boldsymbol{P})$为字符向量,$\boldsymbol{f}(\boldsymbol{P})$为附加特征向量,[,]为连接操作.

字符向量可以包括 OOV 词汇,附加特征向量可以提供语义和语法特征,这些特征不被包括在预训练的词向量中. 因此,模型使用以上四种向量的连接作为最终嵌入可以获得更多信息并带来更好的识别效果.

3.2　基于注意力机制的上下文双向长短期记忆网络编码器构建

编码层对句子的上下文信息进行编码,HPIN 使用基于注意力机制的上下文双向长短期记忆网络作为编码器. 双向长短期记忆网络包括两个方向相反的长短期记忆网络,能够学习句子的前向和逆向两个方向的上下文信息. 上下文长短期记忆(Contextual Long Short-Term Memory Network, Contextual-LSTM)网络不是仅使用 LSTM 的最后一个单元的输出,而是使用所有单元的隐藏状态作为输出,获得 LSTM 上的所有单元的信息.

本文设计的 ABC-BiLSTM 结合了 BiLSTM 和

Contextual-LSTM 的优点,能够获取前向 LSTM 和逆向 LSTM 所有单元的隐藏状态,并且在此基础上加入注意力机制,为不同单元的隐藏状态对句子编码结果的影响提供权重信息,从而产生更好的编码性能. 其工作原理如下.

对于长度为 l 的句子的词向量矩阵 $\boldsymbol{w}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_l)$,编码过程为:

$$\overrightarrow{\boldsymbol{h}_i}=\overrightarrow{L}(\boldsymbol{w}_i,\overrightarrow{\boldsymbol{h}_{i-1}}), \tag{2}$$

$$\overleftarrow{\boldsymbol{h}_i}=\overleftarrow{L}(\boldsymbol{w}_i,\overleftarrow{\boldsymbol{h}_{i+1}}), \tag{3}$$

$$\boldsymbol{h}_i=[\overrightarrow{\boldsymbol{h}_i},\overleftarrow{\boldsymbol{h}_i}], \tag{4}$$

$$\boldsymbol{C}(\boldsymbol{w})=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_l]. \tag{5}$$

式中:$\boldsymbol{w}_i$ 为 LSTM 第 i 节点的输入,$\overrightarrow{L}$ 为前向 LSTM,$\overleftarrow{L}$ 为逆向 LSTM,$\overrightarrow{\boldsymbol{h}_i}$为 $\overrightarrow{L}$ 第 i 节点的隐藏状态,$\overleftarrow{\boldsymbol{h}_i}$为 $\overleftarrow{L}$ 第 i 节点的隐藏状态,$\boldsymbol{C}(\boldsymbol{w})$ 为 Contextual-LSTM 的输出.

增加的注意力机制为:

$$\boldsymbol{A}_{ij}=\boldsymbol{w}_\alpha^T[\boldsymbol{h}_i,\boldsymbol{h}_j,\boldsymbol{h}_i\otimes\boldsymbol{h}_j], \tag{6}$$

$$\boldsymbol{a}(\boldsymbol{h}_i)=\sum_{i=1}^{m}\frac{\exp(\boldsymbol{A}_{ij})}{\sum_{k=1}^{m}\exp(\boldsymbol{A}_{kj})}\boldsymbol{h}_j, \tag{7}$$

$$A(\boldsymbol{w})=[\boldsymbol{a}(\boldsymbol{h}_1),\boldsymbol{a}(\boldsymbol{h}_2),\cdots,\boldsymbol{a}(\boldsymbol{h}_l)]. \tag{8}$$

式中:$\boldsymbol{w}_\alpha\in\boldsymbol{R}^{3\mathrm{d}}$是可训练的参数,⊗为元素按位相乘操作,[,]为连接操作,$\boldsymbol{A}(\boldsymbol{w})$ 为 ABC-BiLSTM 的输出,$i,j\in[1,\cdots,l]$.

由此可见,注意力机制的增加改变了 Contextual-LSTM 隐藏层节点状态 h_i 对于编码结果中每一列影响的权重,由于注意力机制中的参数 $\boldsymbol{w}_\alpha$ 是可训练的参数,可通过选择合适的损失函数,训练 $\boldsymbol{w}_\alpha$,获得更好的编码结果.

3.3 基于多种矩阵运算的匹配操作

匹配层对来自编码层的句子编码矩阵对进行匹配. 与以往单纯的将句子编码矩阵相乘或者相减作为匹配结果不同,HPIN 对编码层输出的矩阵对进行多种矩阵计算,包括矩阵相减、矩阵相减后按位取绝对值、矩阵按位乘法,其目的是获取编码矩阵对之间的相关性,最后把句子编码矩阵对和 3 种匹配结果矩阵堆叠起来,形成的三维张量作为最终的匹配结果,如下

$$\boldsymbol{m}=\{\boldsymbol{u},\boldsymbol{v},\boldsymbol{u}-\boldsymbol{v},|\boldsymbol{u}-\boldsymbol{v}|,\boldsymbol{u}\otimes\boldsymbol{v}\}. \tag{9}$$

式中:$\boldsymbol{u}$ 和 $\boldsymbol{v}$ 表示两个句子的编码结果矩阵,操作符 |-| 和⊗都是矩阵按位(element-wise)操作,- 为矩阵减法,|-| 为矩阵相减以后按位取绝对值,⊗为按位相乘,{,} 为堆叠操作,即把二维张量堆叠为三维张量,$\boldsymbol{m}$ 为匹配层的匹配结果.

3.4 Xception 特征提取器构建

传统的"编码-匹配"模式难以从匹配结果中提取到足够的分类信息,因此本文设计了"编码-匹配-提取"架构,添加了特征提取层. 在传统的"编码-匹配"模型中,由于缺少提取分类信息的结构,匹配层的匹配结果被直接输入到输出层用于分类,导致了分类准确率下降. 在 HPIN 中,增加特征提取层,用于从匹配结果更好地提取分类信息. 根据"特征提取器选择实验及分析"实验结果,最终选取 Xception 作为特征提取器.

Xception 是 Chollet 于 2017 年提出的深度学习模型[13],最早用于图像分类. Xception 是对 Inception 的改进,Chollet 将 Inception 中的 Inception 单元替换为深度可分离卷积单元,得到了 Xception. Xception 是带有残差连接的深度可分离卷积单元的线性堆叠. 简化的深度可分离卷积单元的结构见图 2.

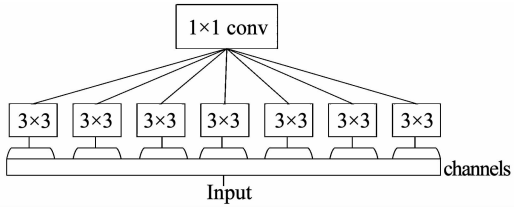

图 2 深度可分离卷积单元

Fig. 2 Depthwise separable convolution module

从图 2 中可以看出,输入通过多个 3×3 卷积核进行卷积,然后结果被连接起来,再进行 1×1 卷积. Xception 由 34 个类似结构的深度可分离卷积单元组成.

在 HPIN 中,Xception 接受来自匹配层的匹配结果(一个三维张量)作为输入,输入数据依次通过多个深度可分离卷积单元,在每个深度可分离卷积单元中,输入数据,先按照通道分组,对每个通道做一次 3×3 的卷积,然后再对卷积结果进行 1×1 的卷积. 深度可分离卷积保证了得到的特征之间独立性,没有太多的相互依赖. 残差连接把一些深度可分离卷积单元之间连接起来,从而避免了梯度爆炸问题. 最后,整个 Xception 的输出被平铺成一个长向量,作为输出层的输入.

由于增加 Xception 作为特征提取器,利用其结构中多个深度可分离卷积单元以及残差连接,有效地提取了句子的分类信息,从而使得模型的分类准确率有了进一步提升.

4 实验设置

4.1 数据集

1) Quora 问题对数据集. Quora 问题对数据集来

源于 Quora. com，包含超过 40 万对真实数据. 每个问题对都有二进制注释，1 表示重复（释义对），0 表示不重复（非释义对）.

2）Twitter Paraphrase SemEval 2015 数据集. 最近不少释义识别研究都采用了 Twitter Paraphrase SemEval 2015 提供的数据集[14]（以下简称 PIT 数据集）. 该数据集由带有噪音的短文本组成，共有 18 762 个文本对.

表 1　数据集划分

Tab. 1　Datasets partition

数据集	总量	训练集	验证集	测试集
Quora	404 290	384 290	10 000	10 000
PIT	18 762	13 063	4 727	972

在所有实验中，数据集划分为训练集、验证集和测试集，如表 1 所示. 对于 Quora 数据集，随机抽取 10 000 个数据作为验证集，10 000 个作为测试集，其余数据作为训练集. 对于 PIT 数据集，使用数据集本身提供的数据划分.

4.2　评价指标

实验采用准确率和 F1 值作为评价指标. 准确率是正确分类的释义对的百分比，F1 值是精确度和召回率的组合. 在使用 Quora 数据集的实验中使用准确率，在使用 PIT 数据集的实验中使用 F1 值，以更好地与其他人的工作进行对比（Quora 数据集中更常用准确率，PIT 数据集中更常用 F1 值）.

4.3　通用设置

实验使用 Keras 框架实现提出的模型，使用初始学习率为 0. 001 的 RMSProp 优化器优化可训练的参数. 批量大小设置为 128. 使用 300 维 840B 语料训练的 GloVe 向量作为预训练词向量. 设置句子标准长度为 32，超出部分会被截去，不足部分用 0 补齐. 对于所有实验，选择在验证集上表现最佳的模型，然后在测试集上对其进行评估.

5　实验结果和分析

5.1　附加特征筛选实验及分析

HPIN 的嵌入层中使用了附加特征向量. 本小节希望探索哪些附加组合可以更好地优化模型效果，并评估附加特征的优化效果在不同规模的数据集上的表现.

5.1.1　单附加特征筛选实验与分析

本实验目的是探讨哪些附加特征能优化模型效果. 由于实验只用于评价单个附加特征对模型的影响，不作模型性能评估，所以仅使用 Quora 数据集.

在该实验中，将备选特征分别添加到模型中，评估该特征的加入对模型准确率的影响. 将没有任何特征添加的模型视为该实验的基线. 实验结果见表 2，其中句子的长度、句子中单词的位置和 n-gram 重叠特征对提高模型准确率没有帮助，而 BTM 特征、词性标注和 Wordnet 相似度的加入提高了模型的准确率.

表 2　单个特征对模型的影响

Tab. 2　Impact of single features on the model

特征	准确率/%
无特征添加	87. 35
句子长度	86. 85
单词位置	87. 14
n-gram 重叠	87. 29
BTM 特征	87. 56
词性标注	88. 37
Wordnet 相似度	88. 48

句子的长度、单词位置和 n-gram 重叠在被添加到词向量中时会产生负面效应. 原因可能是这些特征包含的信息不足，而当它们被添加到词向量中时，同时也将噪声带入了词向量.

5.1.2　附加特征组合筛选实验及分析

本实验的目的是探索哪些附加组合可以更好地优化模型效果. 由于实验只用于评价附加特征组合对模型的影响，不作模型性能评估，所以仅使用 Quora 数据集.

对“单附加特征评估实验与分析”中能优化模型效果的三个特征：BTM 特征、词性标注和 Wordnet 相似度，进行组合并通过实验对这些组合的效果进行评估，结果如表 3 所示. 可以发现“词性标注 + Wordnet 相似度”效果更好，因此模型最终选择词性标注和 Wordnet 相似度的组合生成附加特征向量.

由表 3 可知，BTM 特征在单独添加到词向量中时会产生正面影响，而当它被添加到具有 Wordnet 相似度或词性标注的词向量中时，模型表现并不好. 原因可能是 BTM 特征携带的信息与 Wordnet 相似度以及词性标注携带的信息存在重叠，当同时被加入模型中时，噪声比有价值的信息增加得更多.

5.1.3　附加特征在不同数据集上对模型优化效果评估

本组实验的目的是评估附加特征对于所提出模型 HPIN 的优化效果在不同规模数据集上的表现. 主要记录 4 组结果：无附加特征添加的模型分别在 Quora（大型数据集）和 PIT（中小型数据集）上的准确率和有“词性标注 + Wordnet 相似度”作为附加特征添加的模型分别在 Quora 和 PIT 数据集上的准确率. 实验结果如表 4 所示.

表 3 附加特征组合对模型的影响

Tab. 3 Impact of feature combinations on the model

组合	准确率/%
无特征添加	87.35
BTM 特征	87.56
词性标注	88.37
Wordnet 相似度	88.48
BTM 特征 + 词性标注 + Wordnet 相似度	88.57
BTM 特征 + 词性标注	88.23
BTM 特征 + Wordnet 相似度	88.18
词性标注 + Wordnet 相似度	88.58

表 4 附加特征对模型影响(Quora,PIT)

Tab. 4 Impact of additional features on the model(Quora,PIT)

附加特征	准确率/%(Quora)	准确率/%(PIT)
无附加特征	87.35	83.89
词性标注 + Wordnet 相似度	88.58	86.51

由表 4 可知,对于 Quora 数据集,附加特征的添加使得模型准确率提升了 1.23%. 而对于 PIT 数据集,附加特征的添加使得准确率提升了 2.62%. 显然,附加特征对于模型的优化效果在中小型数据集上表现得更明显.

5.2 特征提取器选择实验及分析

设计特征提取层是为了从匹配结果中更好地提取分类信息. 本实验的目的是验证特征提取层的有效性以及寻找适合的特征提取器. 由于该实验只评估不同特征提取器对模型准确率的影响,不作模型性能评估,所以仅使用 Quora 数据集.

在其他设置不变的情况下,只改变特征提取层的结构,以评估不同特征提取器对模型准确率的影响. 其中,无特征提取层的模型作为实验的基线. 参与对比实验的特征提取器结构有 InceptionV3、DenseNet121、DenseNet169、DenseNet201、Xception、InceptionResnetV2 和 ResNet50. 实验结果如表 5 所示,最佳结果在表格中用下划线标出. 显然有特征提取层的模型比无特征提取层的模型准确率更高. 这表明了特征提取层的设置是有效的. 在各种特征提取器中,Xception 和 DenseNet121 表现最好,达到了 88.5% 以上的准确率. 而 Xception 比 DenseNet121 参数更少,训练得更快,所以最终选择了 Xception 作为模型的特征提取器.

在表 5 中,可以发现 Xception 的性能优于 Inception、DenseNet 和 Resnet. Xception 比 InceptionV3 更深却与 InceptionV3 的参数数量几乎相同,这体现了 Xception 能更有效地使用模型参数. Resnet50 和 DenseNet 结构表现不佳的原因可能是当数据集较小时这些结构更容易过拟合.

表 5 特征提取器效果评估

Tab. 5 Feature extractor effect evaluation

特征提取器	准确率/%
无特征提取层	87.23
InceptionV3	88.22
DenseNet121	88.53
DenseNet169	88.47
DenseNet201	87.45
Xception	<u>88.58</u>
InceptionResnetV2	88.31
ResNet50	88.14

5.3 模型性能评估

本组实验对 HPIN 与其他释义识别模型在 Quora 数据集和 PIT 数据集(分别作为大型数据集和中小型数据集的代表)的释义识别结果进行评估对比.

5.3.1 Quora 数据集上的模型性能评估

将 HPIN 与 GenSen[15]、BiMPM[10]、SSE[17]、ESIM[11]、inferSent[18]和 PWIM[20]在 Quora 数据集上释义识别的结果进行比较. 与 HPIN 对比的模型数据来源于文献[10,15,21]. 结果如表 6 所示,最佳结果在表格中用下划线标出. HPIN 在测试集上达到了 88.58% 的准确率,这比 BiMPM 的 88.17% 表现得更好.

表 6 Quora 数据集上的模型评估

Tab. 6 Model evaluation on Quora

模型	准确率/%
GenSen	87.01
BiMPM	88.17
SSE	87.80
ESIM	85.40
inferSent	86.60
PWIM	83.40
HPIN	<u>88.58</u>

HPIN 表现得比较好的原因可能有三点:首先是附加特征的使用,参与对比的其他模型没有使用附加特征,而 HPIN 采用多特征融合词向量,其蕴含的信息比普通的预训练词向量更加丰富;其次是增设了特征提取层,参与对比的其他模型没有特征提取步骤,HPIN 使用 Xception 作为特征提取器,而“特征提取器选择实验及分析”表明,Xception 作为特征提取器能够进一步提取分类信息,从而提升了模型识别的准确率;最后是 HPIN 在编码层使用了注意力机制,该机制能调节不同隐藏层状态对编码结果

影响的权重，从而有助于准确率的提升，而其他几个模型没有使用注意力机制.

5.3.2　PIT 数据集上的模型性能评估

将 HPIN 与 Huang 等的模型[16]、AugDeepParaphrase 模型[1]、SSE[17]、ESIM[11]、inferSent[18]和 PWIM[20]在 PIT 数据集上释义识别的结果进行比较. 与 HPIN 对比的模型数据来源于文献[1,16,21]. 结果如表 7 所示，最佳结果在表格中用下划线标出. HPIN 的 $F1$ 值为 0.749，仅比最佳模型 AugDeepParaphrase 低 0.002. 表明 HPIN 不仅在 Quora 这样的大型数据集上表现良好(见表 6)，在像 PIT 这样的中小型数据集上也有很好的表现.

表 7　PIT 数据集上的模型评估

Tab.7　Model evaluation on PIT

模型	$F1$ 值
Huang 等	0.650
AugDeepParaphrase	<u>0.751</u>
SSE	0.422
ESIM	0.538
inferSent	0.451
PWIM	0.656
HPIN	0.749

同时，分析表 6 和表 7 结果可知，模型 SSE、ESIM、inferSent 和 PWIM 在大型数据集 Quora 上表现良好(准确率与最优模型差距不大)，但在中小型数据集 PIT 上则表现较差(表现远远差于最优模型)，表明了这些模型对数据集大小较敏感；而 HPIN 在大型数据集 Quora 和中小型数据集 PIT 上都取得了良好的效果，表明 HPIN 具有一定程度的泛用性. 其原因一是模型采用了多特征融合的词向量，特别是当数据集较小时，附加特征的贡献尤为明显(结论来自“附加特征在不同数据集上对模型优化效果评估”)；另一个原因则是增设了特征提取层，充分提取了匹配结果中的分类信息，无论在大型数据集上还是在中小型数据集上都具有良好效果.

6　结　论

本文构建了一种新的释义识别模型 HPIN. 与大多数现有的释义识别模型采用的“编码-匹配”模式不同，采用“编码-匹配-提取”模式，增设了特征提取层，从匹配结果中提取更深层的分类信息. HPIN 是一个分层模型，由 6 层组成：输入层、嵌入层、编码层、匹配层、特征提取层、输出层. 嵌入层使用可训练的词向量、不可训练的词向量、字符向量和附加特征向量的连接，作为最终的词向量，较普通的预训练词向量携带更丰富的信息；编码层中采取基于注意力机制的上下文双向 BiLSTM 作为编码器，获取前向和逆向两个 LSTM 中所有隐藏层中的信息，有效地对词向量矩阵的上下文进行编码；在匹配层中，运用多种矩阵运算，从不同角度获取句子对的匹配信息；在特征提取层中，使用 Xception 结构，更有效地提取分类信息. 本文在 Quora(作为大型数据集代表)和 PIT 两个公开数据集上(作为中小型数据集的代表)评估该模型，均达到了竞争性的效果，从而表明所提出的 HPIN 模型不仅能有效提高释义识别的准确率，而且在不同规模的数据集上(Quora 和 PIT)都表现良好，因此也具有一定程度的泛用性.

参考文献

[1] AGARWAL B, RAMAMPIARO H, LANGSETH H, et al. A deep network model for paraphrase detection in short text messages[J]. Information Processing & Management, 2018, 54(6): 922. DOI: 10.1016/j.ipm.2018.06.005

[2] YIN Wenpeng, SCHÜTZE H. Convolutional neural network for paraphrase identification[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2015: 901. DOI: 10.3115/v1/N15-1091

[3] MADNANI N, TETREAULT J, CHODOROW M. Re-examining machine translation metrics for paraphrase identification[C]// Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2012: 182. DOI: 10.1007/978-3-319-49130-1_33

[4] DAS D, SMITH N A. Paraphrase identification as probabilistic quasi-synchronous recognition [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. Stroudsburg, USA: Association for Computational Linguistics, 2009: 468

[5] SAHI M, GUPTA V. A novel technique for detecting plagiarism in documents exploiting information sources[J]. Cognitive Computation, 2017, 9(6): 852. DOI: 10.1007/s12559-017-9502-4

[6] VANI K, GUPTA D. Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparisons, analysis and challenges[J]. Information Processing and Management, 2018, 54(3): 4082. DOI: 10.1016/j.ipm.2018.01.008

[7] JIANG Yuncheng, BAI Wen, ZHANG Xiapei, et al. Wikipedia-based information content and semantic similarity computation[J]. Information Processing and Management, 2017, 53(1): 248. DOI: 10.1016/j.ipm.2016.09.001

[8] FRANCO-SALVADOR M, ROSSO P, MONTES-Y-GÓMEZ M. A systematic study of knowledge graph analysis for cross-language plagiarism detection[J]. Information Processing and Management, 2016, 52(4): 550. DOI: 10.1016/j.ipm.2015.12.004

[9] 黄江平，姬东鸿. 基于卷积网络的句子语义相似性模型[J]. 华南理工大学学报(自然科学版)，2017，45(3)：68
HUANG Jiangping, JI Donghong. Sentence semantic similarity model

based on convolutional network[J]. Journal of South China University of Technology (Natural Science Edition), 2017, 45(3): 68. DOI: 10.3969/j.issn.1000-565X.2017.03.010

[10] WANG Z, HAMZA W, FLORIAN R. Bilateral multi-perspective matching for natural language sentences[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Marina del Rey, CA: IJCAI, 2017: 4144. DOI: 10.24963/ijcai.2017/579

[11] CHEN Qian, ZHU Xiaochen, LING Zhenhua, et al. Enhanced LSTM for natural language inference[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2017: 1657. DOI: 10.18653/v1/P17-1152

[12] KIM S, KANG I, KWAK N. Semantic sentence matching with densely-connected recurrent and co-attentive information[C]// Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019. DOI: 10.1609/aaai.v33i01.33016586

[13] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii: CVPR, 2017: 1251. DOI: 10.1109/CVPR.2017.195

[14] XU W, CALLISON-BURCH C, DOLAN B. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT)[C]// Proceedings of the 9th International Workshop on Semantic Evaluation. Colorado: Association for Computational Linguistics, 2015: 1. DOI: 10.18653/v1/S15-2001

[15] SUBRAMANIAN S, TRISCHLER A, BENGIO Y, et al. Learning general purpose distributed sentence representations via large scale multi-task learning[C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018

[16] HUANG Jiangping, YAO Shuxin, LÜ Chen, et al. Multi-granularity neural sentence model for measuring short text similarity[C]//Proceedings of the International Conference on Database Systems for Advanced Applications. Cham: Springer, 2017: 439. DOI: 10.1007/978-3-319-55753-3_28

[17] NIE Y, MOHIT B. Shortcut-stacked sentence encoders for multi-domain inference[C]//Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. Stroudsburg, USA: Association for Computational Linguistics, 2017: 41. DOI: 10.18653/v1/W17-5308

[18] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2017: 670. DOI: 10.18653/v1/D17-1070

[19] TOMAR G S, DUQUE T, TÄCKSTRÖM O, et al. Neural paraphrase identification of questions with noisy pretraining[C]// Proceedings of the 1st Workshop on Subword and Character Level Models in NLP. Stroudsburg, USA: Association for Computational Linguistics, 2017: 142. DOI: 10.18653/v1/W17-4121

[20] HE H, JIMMY L. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA: Association for Computational Linguistics, 2016: 937. DOI: 10.18653/v1/N16-1108

[21] LAN Wuwei, XU Wei. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering[C]//Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2018: 3890

(编辑 苗秀芝)

(上接第 151 页)

[8] 于烨, 黄默, 王小青, 等. 利用最小一乘法改进的灰色模型的导航卫星钟差预报[J]. 测绘通报, 2019(4): 1
YU Ye, HUANG Mo, WANG Xiaoqing, et al. Navigation satellite clock bias prediction based on grey model improved by least absolute deviations[J]. Bulletin of Surveying and Mapping, 2019(4): 1. DOI: 10.13474/j.cnki.11-2246.2019.0102

[9] 李成龙, 陈西宏, 刘继业, 等. 利用自适应 TS-IPSO 优化的灰色系统预报卫星钟差[J]. 武汉大学学报(信息科学版), 2018, 43(6): 854
LI Chenglong, CHEN Xihong, LIU Jiye, et al. Predicting satellite clock errors using grey model optimized by adaptive TS-IPSO[J]. Geomatics and Information Science of Wuhan University, 2018, 43(6): 854. DOI: 10.13203/j.whugis20160101

[10] 石宁, 卢辰龙, 杨登科, 等. 基于拉格朗日插值的灰色模型卫星钟差预报[J]. 测绘技术装备, 2018, 20(4): 5
SHI Ning, LU Chenlong, YANG Dengke, et al. Grey model satellite clock error prediction based on Lagrange interpolation[J]. Geomatics Technology and Equipment, 2018, 20(4): 5. DOI: 10.3969/j.issn.1674-4950.2018.04.002

[11] 路晓峰, 杨志强, 贾小林, 等. 灰色系统理论的优化方法及其在卫星钟差预报中的应用[J]. 武汉大学学报(信息科学版), 2008, 33(5): 492
LU Xiaofeng, YANG Zhiqiang, JIA Xiaolin, et al. Parameter optimization method of gray system theory for the satellite clock error predicating[J]. Geomatics and Information Science of Wuhan University, 2008, 33(5): 492

[12] 蔡成林, 何成文, 韦照川. 一种 GPS IIR-M 型卫星超快星历钟差预报的高精度修正方法[J]. 测绘学报, 2016, 45(7): 782
CAI Chenglin, HE Chengwen, WEI Zhaochuan. A high-precision correction method of ultra-rapid ephemeris clock bias prediction for GPS block IIR-M satellites[J]. Acta Geodaetica et Cartographica Sinica, 2016, 45(7): 782. DOI: 10.11947/j.AGCS.2016.20160017

[13] 于烨, 张慧君, 李孝辉. 含误差预报校正的 GM(1,1)卫星钟差预报新方法[J]. 测绘科学, 2019, 44(4): 8
YU Ye, ZHANG Huijun, LI Xiaohui. A new method of GM(1,1) satellite clock bias prediction with error prediction correction[J]. Science of Surveying and Mapping, 2019, 44(4): 8. DOI: 10.16251/j.cnki.1009-2307.2019.04.002

[14] 韩晓东, 贺兆礼. 灰色 GM(1,1)与线性回归组合模型及其在变形预测中的应用[J]. 淮南矿业学院学报, 1997, 17(4): 51
HAN Xiaodong, HE Zhaoli. Combination model of GM(1,1) and linear regression and its application in deformation prediction[J]. Journal of Huainan Mining Institute, 1997, 17(4): 51

[15] 刘强, 孙际哲, 陈西宏, 等. CPSO-LSSVM 在自回归钟差预报中的应用[J]. 吉林大学学报(工学版), 2014, 44(3): 807
LIU Qiang, SUN Jizhe, CHEN Xihong, et al. Application analysis of CPSO-LSSVM algorithm in AR clock error prediction[J]. Journal of Jilin University (Engineering and Technology Edition), 2014, 44(3): 807. DOI: 10.13229/j.cnki.jdxbgxb201403036

(编辑 苗秀芝)