融合语义角色和自注意力机制的中文文本蕴含识别

张志昌,曾扬扬,庞雅丽
(西北师范大学计算机科学与工程学院,甘肃兰州 730000)

摘 要: 文本蕴含识别旨在识别两个给定句子之间的逻辑关系.本文通过构造语义角色和自注意力机制融合模块,把句子的深层语义信息与 Transformer 模型的编码部分相结合,从而增强自注意力机制捕获句子语义的能力.针对中文文本蕴含识别在数据集上存在规模小和噪声大的问题,使用大规模预训练语言模型能够提升模型在小规模数据集上的识别性能.实验结果表明,提出的方法在第十七届中国计算语言学大会中文文本蕴含识别评测数据集 CNLI 上的准确率达到了 80.28%.

关键词: 自然语言处理;文本蕴含;自注意力机制;语义角色标注;预训练语言模型
中图分类号: TP391.1;TP183 **文献标识码:** A **文章编号:** 0372-2112 (2020)11-2162-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.11.010

A Chinese Textual Entailment Recognition Method Incorporating Semantic Role and Self-Attention

ZHANG Zhi-chang, ZENG Yang-yang, PANG Ya-li
(*College of Computer Science and Engineering, Northwest Normal University, Lanzhou, Gansu* 730000, *China*)

Abstract: Recognizing textual entailment is intended to infer the logical relationship between two given sentences. In this paper, we incorporate the deep semantic information of sentences and the encoder of Transformer by constructing the SRL-Attention fusion module, and it effectively improves the ability of self-attention mechanism to capture sentence semantics. Furthermore, concerning the small scale and high noise problems on the dataset, we use large-scale pre-trained language model improving the recognition performance of model on small-scale dataset. Experimental results show that the accuracy of our model on the dataset CNLI, it is released as Chinese textual entailment recognition evaluation corpus at the 17th China National Conference on Computational Linguistics, reaches 80.28%.

Key words: natural language processing; textual entailment; self-attention mechanism; semantic role labeling; pre-trained language model

1 引言

文本蕴含用来描述两个文本之间的语义推理关系,这种关系广泛存在于自然语言文本中.文本蕴含具体的定义是:给定两个文本,分别称其为前提句(Premise,P)和假设句(Hypothesis,H),如果人们依据自己的常识认为假设句的语义能够由前提句的语义推理得出,那么称前提句蕴含假设句,记作 $P \Rightarrow H$[1].文本蕴含识别(Recognizing Textual Entailment,RTE)任务不仅要识别两个语句之间是否存在蕴含关系,还要识别矛盾关系和中立关系.表 1 是 CNLI 数据集示例.

表 1 CNLI 数据集示例

蕴含	*P*:五个人在楼梯上排成一行. *H*:一群人站在楼梯上.
矛盾	*P*:女孩在与天空中的彩虹合影. *H*:天空永远不会有彩虹.
中立	*P*:一群人在参加某种跑步比赛. *H*:一群人想要在赛事中赢得奖杯.

从表 1 所示的三组句子可以发现,RTE 涉及到词汇

收稿日期:2019-11-29;修回日期:2020-03-30;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61762081,No.61662067,No.61662068);甘肃省重点研发计划(No.17YF1GA016)

语义、语义推理、社会经验和常识等多个方面的知识,是一项判断计算机是否在一定程度上“理解”文本语义的具有挑战性的研究任务.

当前基于自注意力机制(Self-Attention)的方法被广泛应用在 RTE 模型中[2],但是对于含有复杂语义的中文语句来说,传统的自注意力机制很难完全提取语句的深层特征,为了解决该问题,我们希望人为地提供给模型更多已知的语义特征. 语义角色标注(Semantic Role Labeling,SRL)是浅层语义分析的一种实现方式[3],它能够分析句子中各成分与谓词之间的关系,具备理解句子特定语义的能力,并且已经被证明对广泛的自然语言处理任务是有益的,如机器翻译、问答系统、语篇关系分类等任务[4]. 因此本文试图把 SRL 与中文 RTE 任务相结合,在自注意力机制中融合语义角色来增强自注意力机制捕获句子语义信息的能力.

目前研究者们针对英文文本的 RTE 开展了大量的研究工作,主要是得益于一些高质量大规模数据集的发布. 然而针对中文 RTE 的研究工作仍然较为缺乏. 一方面,受限于语言本身的差异[5],针对英文文本的 RTE 模型通常不能直接应用于中文文本;另一方面,中文的 RTE 语料不多,已有的语料规模偏小. 为了应对数据稀疏的挑战,大规模通用预训练语言模型层出不穷,例如 ELMo[6]、GPT[7]、BERT[8]等. 本文使用在大规模语料上训练的基于全词覆盖的中文 BERT 预训练语言模型(Extend Chinese BERT with Whole Word Masking,BERT-wwm-ext)[9],为语句的编码向量提供更多数据集自身所提取不到的语义信息,尤其对于小规模数据集而言,可以显著提升模型的识别性能.

本文的主要贡献:

(1)本文以神经网络模型 Transformer[10]为基础,在自注意力机制中创新性地融合进语义角色标注信息,提升自注意力机制捕获句子语义的能力.

(2)结合中文文本蕴含识别模型的特点,对句子的语义角色进行提取和编码,为语义角色和自注意力机制的融合提供了可能.

(3)使用大规模预训练语言模型 BERT-wwm-ext,缓解中文文本蕴含识别任务人工标注语料不足导致模型在小规模数据集上泛化能力差的问题.

2　相关工作

文本蕴含识别的主要方法总体可以归纳为两大类:基于传统机器学习的特征工程的方法和基于深度学习的方法.

早期 RTE 主要使用基于特征工程的方法,该方法需要对前提句和假设句进行特征的提取和表示,然后利用各种机器学习方法(如支持向量机、最大熵等)来识别句子之间的关系[11~14]. 因为特征的构造依赖于人工,而且提取到的句子语义信息往往较为浅显,所以基于特征工程的方法对于 RTE 来说远远不够.

近年来基于深度学习的方法在 RTE 研究领域得到了广泛应用[15]. ROCKTASCHE 等人[16]提出 word-by-word attention,该做法可以更好地发现前提句和假设句中词与词之间的关系;WANG 等人[17]在此基础上提出 mLSTM(matching-LSTM)模型,该模型是在 LSTM 的隐藏状态当中拼接注意力权重;ANKUR 等人[2]创新性地把 RTE 问题简单地分解为单词间的对齐问题,构建了 decomposable attention 模型;CHEN 等人[18]则在 decomposable attention 模型的输入和最后的拼接部分做出改进,提出 ESIM(Enhanced LSTM)模型,该模型被广泛用作基准模型.

然而随着 RTE 网络模型的发展,研究者们逐渐发现网络模型提取语义信息的能力依然有限,可以合理引入更多的语句自身特征以及外部知识[19~21]. 本文把语义角色融合到自注意力机制当中,构建了一种新的中文文本蕴含识别方法.

3　中文文本蕴含识别模型

本文提出的融合语义角色和自注意力机制的中文文本蕴含识别模型结构如图 1 所示. 该模型由四个部分组成,分别为语言模型微调模块、SRL 编码模块、SRL-Attention 交互模块、以及分类模块. 模型整体可以看作是一个三类分类器,最终需要判定一对句子的关系类别.

3.1　微调语言模型

本文使用 BERT-wwm-ext 作为预训练语言模型,与 BERT 的主要区别是该模型更改了原预训练阶段的训练样本生成策略,另外 BERT-wwm-ext 不仅使用中文维基百科作为训练语料,而且增加了通用领域的语料数据[9].

3.1.1　编码输入向量

输入向量是由三种特征编码而成,分别为字特征,位置特征以及每个字的标记类别. 特征的编码方式和 Transformer 相同,以拼接的方式把前提句和假设句的特征编码向量聚合在一起. 三种特征编码后的向量可形式化分别表示为式(1)、式(2)和式(3):

$$\boldsymbol{x}^c=\{\boldsymbol{c}_i\}_{i=1}^{L},\boldsymbol{x}^c\in\boldsymbol{R}^{L\times V_1},\boldsymbol{c}_i\in\boldsymbol{R}^{V_1} \tag{1}$$

$$\boldsymbol{x}^l=\{\boldsymbol{l}_i\}_{i=1}^{L},\boldsymbol{x}^l\in\boldsymbol{R}^{L\times L_{\max}},\boldsymbol{l}_i\in\boldsymbol{R}^{L_{\max}} \tag{2}$$

$$\boldsymbol{x}^t=\{\boldsymbol{t}_i\}_{i=1}^{L},\boldsymbol{x}^t\in\boldsymbol{R}^{L\times T},\boldsymbol{t}_i\in\boldsymbol{R}^{T} \tag{3}$$

其中,$\boldsymbol{x}^c$ 表示每对句子以字为基本单位所对应的 one-hot 向量;$\boldsymbol{x}^l$ 表示每对句子字位置信息的 one-hot 向量;$\boldsymbol{x}^t$ 表示每对句子中每个字所对应的标记类别;$\boldsymbol{c}_i$、$\boldsymbol{l}_i$、$\boldsymbol{t}_i$

表示第 i 个字所分别对应的特征值;L 表示前提句和假设句两个句子的最大固定长度;V_1 表示所使用的字典大小;$L_{\max}$ 表示位置信息的最大值;T 表示标记类别个数 2.

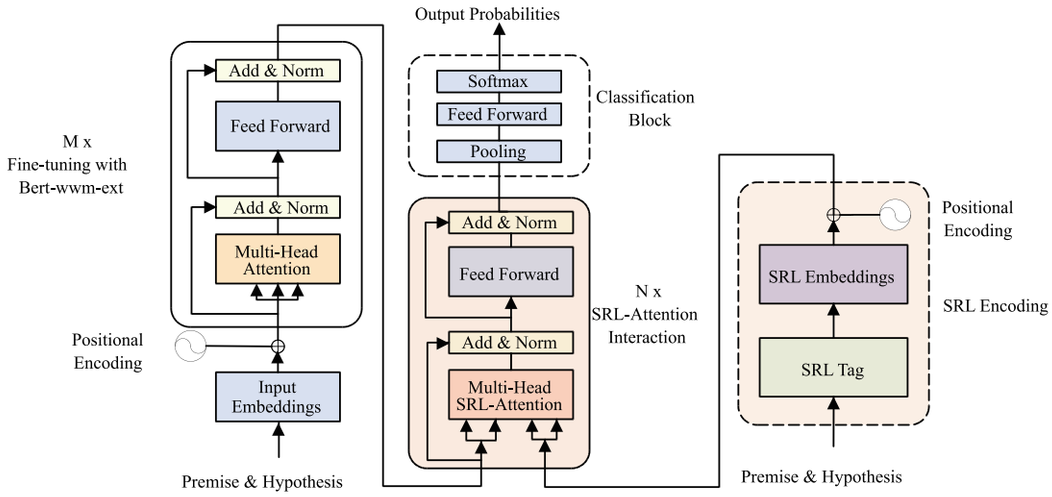


图1 融合语义角色和自注意力机制的中文RTE模型

三种特征向量分别经过不同的投影矩阵 $\boldsymbol{E}_c$、$\boldsymbol{E}_l$、$\boldsymbol{E}_t$,从而得到相同维度大小为 d_c 的特征向量 $\boldsymbol{E}_i^c$、$\boldsymbol{E}_i^l$、$\boldsymbol{E}_i^t$,然后求和得到对应每个字的输入向量 $\boldsymbol{x}_i^e$,编码过程如式(4)~(7)所示,最终的输入向量 $\boldsymbol{x}^e$ 可表示为式(8):

$$\boldsymbol{E}_i^c = \boldsymbol{c}_i\boldsymbol{E}_c, \quad \boldsymbol{E}_c \in \boldsymbol{R}^{V_1 \times d_c} \tag{4}$$

$$\boldsymbol{E}_i^l = \boldsymbol{l}_i\boldsymbol{E}_l, \quad \boldsymbol{E}_l \in \boldsymbol{R}^{L_{\max} \times d_c} \tag{5}$$

$$\boldsymbol{E}_i^t = \boldsymbol{t}_i\boldsymbol{E}_t, \quad \boldsymbol{E}_t \in \boldsymbol{R}^{T \times d_c} \tag{6}$$

$$\boldsymbol{x}_i^e = \boldsymbol{E}_i^c + \boldsymbol{E}_i^l + \boldsymbol{E}_i^t, \quad \boldsymbol{x}_i^e \in \boldsymbol{R}^{d_c} \tag{7}$$

$$\boldsymbol{x}^e = \{\boldsymbol{x}_i^e\}_{i=1}^{L}, \quad \boldsymbol{x}^e \in \boldsymbol{R}^{L \times d_c} \tag{8}$$

3.1.2 BERT-wwm-ext 模块

在本模块中,BERT-wwm-ext 使用 M 个编码块来捕获句子中汉字的局部和全局依赖关系,每个编码块由两个子层组成:多头自注意力层(Multi-Head Attention)和前馈网络层(Feed Forward Network,FFN). 为了解决梯度爆炸和梯度退化问题,每个子层均使用残差连接和层标准化进行堆叠[22].

多头自注意力的主要计算过程:首先对 key 和 query 进行点积运算,然后使用激活函数进行归一化得到注意力权值,最后注意力权值与 value 做加权求和运算,具体的计算过程可以描述为式(9)和式(10):

$$\boldsymbol{x}_{\text{score}}^a = \text{Softmax}(\boldsymbol{k}\boldsymbol{q}^T/\sqrt{d_{h_1}}), \boldsymbol{x}_{\text{score}}^a \in \boldsymbol{R}^{H_1 \times L \times L} \tag{9}$$

$$\boldsymbol{x}^a = \text{Transpose}(\boldsymbol{x}_{\text{score}}^a \boldsymbol{v}), \boldsymbol{x}^a \in \boldsymbol{R}^{L \times d_c} \tag{10}$$

式中的 $\boldsymbol{k}$、$\boldsymbol{q}$、$\boldsymbol{v}$ 是 $\boldsymbol{x}^e$ 进行维度变换后,分别经过三个线性变换所得到,它们的维度均是 $H_1 \times L \times L$,其中,H_1 是多头自注意力的头数目,d_{h_1} 表示矩阵分裂后的维度,即 $d_{h_1} = d_c/H_1$. Transpose 表示矩阵的变换运算,目的是将分裂的多头自注意力集合在一起得到最终的计算结果 $\boldsymbol{x}^a$.

前馈网络主要是由激活函数 ReLU 和两个线性层 Linear 组成,如式(11)~(13)所示:

$$\text{ReLU}(x) = \max(0, x) \tag{11}$$

$$\text{Linear}_k(x) = w_k x + b_k, k \in \{1, 2\} \tag{12}$$

$$\text{FFN}(x) = \text{Linear}_1(\text{ReLU}(\text{Linear}_2(x))) \tag{13}$$

3.2 语义角色编码模块

语义角色编码模块主要包括两个过程:①使用 Pyltp 工具[23]提取语句的浅层语义角色标注信息;②把语义角色编码为特征向量.

本文使用 Pyltp 工具提取数据集中每个句子的 SRL,并使用 BIO 标注模式进行标记. 核心的语义角色有 A_0~A_5六种,A_0通常表示动作的施事,A_1通常表示动作的受事,A_2~A_5根据谓语动词不同会有不同的语义含义,附加语义角色共十六种. SRL 标签示例如表 2 所示.

表 2 SRL 标签示例

P:一名男子在机场擦拭地板.	
H:一名男子目前在机场.	
P:B-A_0 B-A_0 B-A_0 I-A_0 B-LOC B-LOC I-LOC B-V I-V B-A_1 I-A_1 O	
H:B-A_0 B-A_0 B-A_0 I-A_0 B-TMP I-TMP B-V B-A_1 I-A_1 O	

语义角色编码方式与 3.1.1 节中输入向量的编码方式类似,也是由三种特征向量编码而成,它们分别是标签特征,位置特征以及每个标签的标记类别. 首先使用向量嵌入的方式分别得到 $\boldsymbol{E}_i^s$、$\boldsymbol{E}_i^{ls}$、$\boldsymbol{E}_i^{ts}$,然后求和得到对应每个语义角色标签的编码向量 $\boldsymbol{x}_i^b$. $\boldsymbol{x}_i^b$ 可表示为式(14),每对句子的语义角色编码向量 $\boldsymbol{x}^b$ 可表示为式(15):

$$\boldsymbol{x}_i^b = \boldsymbol{E}_i^s + \boldsymbol{E}_i^{ls} + \boldsymbol{E}_i^{ts} \tag{14}$$

$$\boldsymbol{x}^b = \{\boldsymbol{x}_i^b\}_{i=1}^{L}, \boldsymbol{x}^b \in \boldsymbol{R}^{L\times d_s} \tag{15}$$

其中,$\boldsymbol{E}_i^s$、$\boldsymbol{E}_i^{ls}$、$\boldsymbol{E}_i^{ts}$、$\boldsymbol{x}_i^b$ 的维度均为 d_s.

3.3 SRL-Attention 交互模块

SRL-Attention 交互模块的作用是把语义角色的编码信息与 BERT-wwm-ext 所提取的语句特征进行交互融合,整个模块基于 Transformer 的编码框架,创新点在于多头自注意力部分的改进. 与卷积神经网络中使用多个感受野的目的类似,本模块由 N 个交互模块堆叠而成,从不同维度来充分提取特征信息. 每个交互模块是由 Multi-Head SRL-Attention 层和前馈层两个子层组成.

Multi-Head SRL-Attention 层的重点在于:如何根据语义角色标注信息更改注意力权值. 具体过程是分别计算各自的注意力权值,然后进行求和得到更改后的注意力权值,最后将注意力权值与 $\boldsymbol{v}^a$ 做加权求和运算. Multi-Head SRL-Attention 计算过程可以描述为式(16)和式(17):

$$\boldsymbol{x}_{\text{score}}^c = \text{Softmax}(\boldsymbol{k}^a\boldsymbol{q}^{a^T}/\sqrt{d_{h2}}) + \text{Softmax}(\boldsymbol{k}^b\boldsymbol{q}^{b^T}/\sqrt{d_{h3}}),\ \boldsymbol{x}_{\text{score}}^c \in \boldsymbol{R}^{H_2\times L\times L} \tag{16}$$

$$\boldsymbol{x}^c = \text{Transpose}(\boldsymbol{x}_{\text{score}}^c\boldsymbol{v}^a), \boldsymbol{x}^c \in \boldsymbol{R}^{L\times d_c} \tag{17}$$

其中,$\boldsymbol{k}^a$、$\boldsymbol{q}^a$、$\boldsymbol{v}^a$ 是 BERT-wwm-ext 的输出向量分别经过三个线性变化所得,它们的维度均是 $H_2\times L\times d_{h_2}$,$H_2$ 是 Multi-Head SRL-Attention 的头数目,d_{h_2} 表示矩阵分裂后的维度,即 $d_{h_2} = d_c/H_2$. 同理,$\boldsymbol{k}^b$ 和 $\boldsymbol{q}^b$ 是 SRL 编码向量分别经过两个线性变化所得,它们的维度均是 $H_2\times L\times d_{h3}$,即 $d_{h_3} = d_s/H_2$,$\boldsymbol{x}^c$ 是 Multi-Head SRL-Attention 层的最终输出.

3.4 分类模块

分类模块的作用是预测句子对的最终标签,该模块主要包括池化层、FNN 层以及 Softmax 层. 本文使用 BERT-wwm-ext 句子级别的语言模型,在每个句子对的首位置都添加[CLS]标记,然后使用 Transformer 对[CLS]进行深度编码. 由于 Transformer 可以无视空间和距离,从而把全局信息编码进每个位置,因此将[CLS]的输出向量直接作为池化结果,最后依次进入 FNN 层和 Softmax 层,得到每个类别的概率分布. 整个分类模块可形式化描述为式(18):

$$\boldsymbol{y} = \text{Softmax}(\text{FNN}(\boldsymbol{x}_{\text{first}})), \boldsymbol{x}_{\text{first}} \in \boldsymbol{R}^{d_c} \tag{18}$$

4 实验与结果分析

4.1 实验数据

本文使用的实验数据是 CNLI 数据集,是第十七届中国计算语言学大会(The Seventeenth China National Conference on Computational Linguistics,CCL 2018)中文文本蕴含识别评测任务[24]所发布的数据集. 该数据集是在英文文本蕴含数据集 SNLI 和 MultiNLI 的基础上进行人工转译、机器翻译、人工整理等方式综合构建而成. CNLI 数据集统计情况如表 3 所示,数据集在类别上分布比较均匀.

表 3　CNLI 数据集统计表

关系类别	训练集	开发集	测试集
蕴含	29738	3485	3475
矛盾	28937	3417	3343
中立	31325	3098	3182
总计	90000	10000	10000

4.2 评价指标

实验以准确率(Accuracy)作为评价指标,其具体定义如式(19)所示:

$$\text{Accuracy} = \frac{N_c}{N_p}\times 100\% \tag{19}$$

其中,N_c 是关系预测正确的句对数量;N_p 是预测的句对总量.

4.3 参数设置

各个参数的设置情况:模型训练的初始学习率是 2e-5;batch 是 32;语句对的最大固定长度是 115;位置标记最大值是 512;字向量维度是 768;语义角色编码的标签向量维度是 120. 为了控制网络的复杂度,防止过拟合,模型中加入了 dropout 层,在 BERT-wwm-ext 预训练部分,隐层之间以及注意力层中的 dropout 比率初始值均设置为 0.1. 在 SRL-Attention 交互模块隐层之间的 dropout 比率初始值设置为 0.1,SRL-Attention 内部不设置 dropout.

4.4 结果分析

4.4.1 RTE 方法比较

我们选择 8 种基于深度学习的中文 RTE 方法和本文模型进行比较,表 4 展示了不同方法在 CNLI 数据集上的实验结果. 按照模型的组成特点可以将这些方法分为三类:

(1)简单神经网络模型. CNN、BiLSTM 和 CNN + BiLSTM 属于较为简单的神经网络模型,优点是参数量少,但往往准确率不高.

(2)结合注意力机制的方法. Decomposable attention、ESIM、ESIM + Co-attention、attention alignment 这四种方法所代表的是以 RNN、CNN 和注意力机制所结合的一类方法.

(3)基于大规模预训练语言模型的方法,如 ELMo + multiple attention.

表 4 不同方法在 CNLI 数据集上的实验结果

方法	开发集(%)	测试集(%)
CNN	66.31	65.20
BiLSTM	67.98	67.41
CNN + BiLSTM	69.56	68.41
decomposable attention	70.20	69.35
ESIM	73.36	72.22
ESIM + Co-attention	78.01	76.92
attention alignment	78.56	78.28
ELMo + multiple attention	77.42	76.20
SRL-Attention + BERT-wwm-ext	80.54	80.28

本文模型结合了当前主流模型的特点,使用预训练语言模型 BERT-wwm-ext,并把语义角色融合到自注意力机制当中. 由表 4 的实验结果表明,与其他方法相比我们的模型性能表现最优,最终在 CNLI 开发集上的准确率达到了 80.54%,测试集上的准确率达到了 80.28%.

4.4.2 SRL-Attention 模块中层数和头数的选取

SRL-Attention 模块中,层数和头数的选择会直接影响模型的识别精度. 不同层数和头数下开发集的准确率变化情况如图 2 所示,两个参数的总体变化规律是当参数设置越大,模型的性能越好,但是层数设置不能过大,否则准确率反而会下降. 由图 2 可以看出,层数和头数的最优值分别是 4 和 6,此时模型在开发集上的准确率最高可以达到 80.56%.

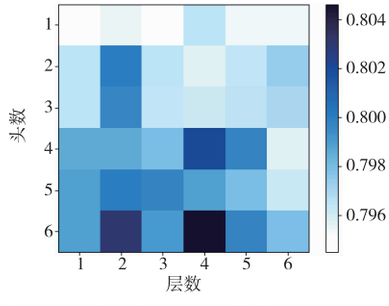


图2 比较不同层数和头数下开发集的准确率

4.4.3 SRL-Attention 模块和 BERT-wwm-ext 的有效性

为证明 SRL-Attention 模块和 BERT-wwm-ext 的有效性,使用四个不同语言模型 GloVe、BERT、BERT-wwm 和 BERT-wwm-ext 分别进行四组对比实验. 实验结果如表 5 所示. 当四个语言模型下游不使用 SRL-Attention 模块时,由于 BERT-wwm-ext 使用了全词掩码的方法,并且预训练语料规模比 BERT-wwm 大,因此获得了最佳的性能. 使用 SRL-Attention 模块的网络模型在开发集和测试集上的准确率也都有不同程度地提高. 由此说明 SRL-Attention 模块和 BERT-wwm-ext 在我们的模型中都具有关键性的作用.

表 5 SRL-Attention 模块和 BERT-wwm-ext 有效性的证明结果

语言模型	无 SRL-Attention(%)		有 SRL-Attention(%)	
	开发集	测试集	开发集	测试集
GloVe	72.36	71.47	75.82	75.30
BERT	78.90	78.87	79.29	79.13
BERT-wwm	79.0	79.03	79.42	79.32
BERT-wwm-ext	79.22	79.25	80.54	80.28

本文模型中融合了语义角色标注信息,本质上是进一步地给模型提供了语言的语义知识. 我们在不同规模大小的 CNLI 训练集上分别训练出两个模型,两个模型分别表示是否使用 SRL-Attention 模块,然后都在测试集和开发集上进行测试. 不同规模大小的训练集所得到的测试结果如图 3 所示. 当训练数据量为 1 万对时,我们的模型在开发集和测试集上的识别准确率都大幅度提高,分别达到了 78.96% 和 78.56%. 因此,语义角色和自注意力机制相融合的方法能够显著提升在小规模数据集上模型的识别性能.

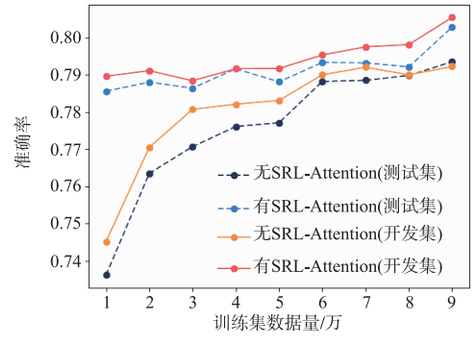


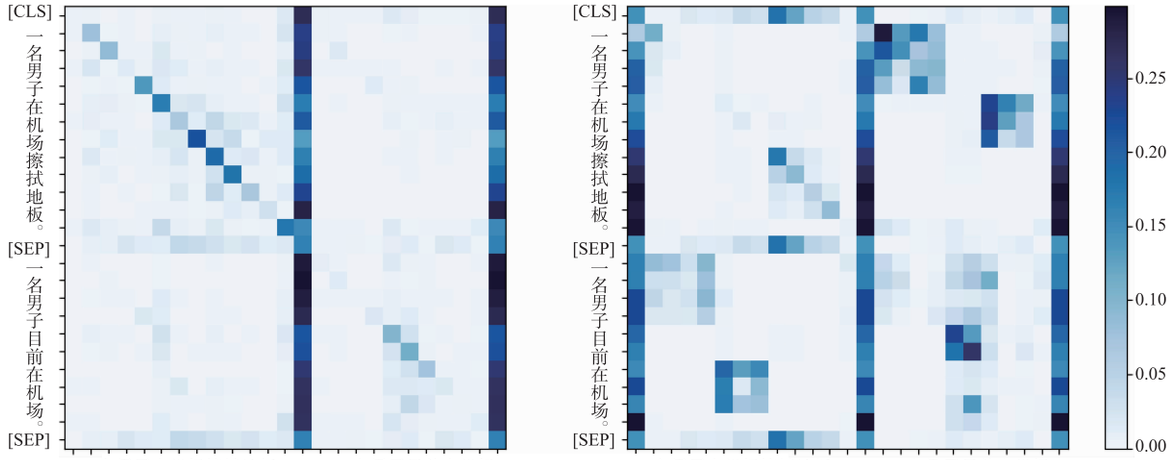
图3 不同规模大小的训练集所得到的测试结果

4.4.4 注意力权值可视化

本文所提出的中文 RTE 网络模型多处用到了注意力机制,注意力机制可以抽象地理解为神经网络从不同局部学习语义的重要性. 注意力机制的权值大小就反映了两个字之间的关系程度以及相似度. 因此可以通过可视化权值矩阵,更加直观地观察语义角色对注意力权值的影响.

以表 2 中的示例作为可视化的对象,当网络模型不使用 SRL-Attention 模块时,如果只使用语言模型 BERT-wwm-ext,两个句子会被错误地判断为中立关系. 图 4 是注意力权值矩阵对比热力图,所使用的权值矩阵都是 BERT-wwm-ext 模块中第 12 层多头自注意力的第 12 个头的注意力权值矩阵,图 4(*a*)是没有 SRL-Attention 模块的网络模型在预测关系时所产生的权值矩阵热力图,图 4(*b*)是使用了 SRL-Attention 模块的完整网络模型在预测关系时所产生的权值矩阵热力图. 在

热力图中,颜色越深表示两个字之间的关系越紧密. 在对比两个热力图的不同点时可以明显地发现:图 4(*b*)中的“在”、“机”、“场”所对应的权值都比较大,而这些字恰恰是判断表 2 示例是蕴含关系的重要特征. 由此可以直观地说明:在自注意力机制中融合语义角色标注信息,可以提升自注意力机制捕获句子语义的能力.



（*a*）无SRL-Attention　　（*b*）有SRL-Attention

图4　注意力权值矩阵对比热力图

4.4.5　预测错误原因分析

我们使用本文提出的模型和基于 BERT-wwm-ext 的模型分别在 CNLI 的测试集上进行预测,图 5 是预测错误的结果统计图. 在两个预测结果当中存在 1576 对相同的句子. 图 5 中的第二列和第三列表示:在去除 1576 对相同的句子之后,两种方法所分别对应的预测错误的数据分布情况.

表 6 是预测错误的数据集示例. 结合图 5 和表 6,对于错误原因进行以下分析:

表 6　预测错误的数据集示例

序号	前提句	假设句	正确关系	预测结果
1	只有庭院以及巨大的门保存了下来,其余的部分已被重建了.	庭院和门户,是结构中唯一没有被重建的部分.	蕴含	矛盾
2	我看到那些用火和血铸造民族的国家,他们创造了一个美丽的梦想.	看到这个国家的诞生令人难以置信.	中立	蕴含
3	一个穿着橙色衬衫的年轻男孩看着什么.	男孩正在看电视.	中立	矛盾

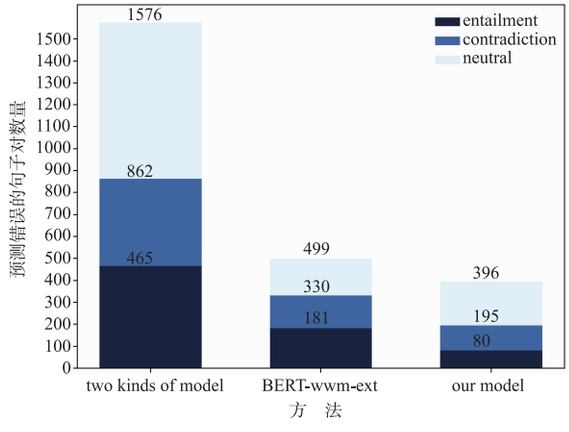


图5　预测错误的结果统计图

(1)数据集存在噪音. 因为 CNLI 数据集是在英文数据集的基础上以人工转译、机器翻译、人工整理等方式构建而成的,某些句子在经过翻译之后,甚至会改变句子之间的逻辑关系,所以数据集存在噪音在一定程度上会对模型造成干扰.

(2)部分句子的语义较为复杂,推理过程涉及到了深层的逻辑推理,常识性推理等. 如图 5 所示,存在 1576 对句子使得两种模型都未能正确预测,这些句子很多涉及到了深层的逻辑推理,常识性推理等. 如表 6 中 1 号句子对,前提句和假设句之间存在较强的逻辑推理关系,神经网络模型只能根据假设句中的否定词“没有”,最终错误地把关系预测为矛盾. 表 6 中 2 号句子对则涉及到了常识性推理,需要基于人类的先验知识进行推理,这是文本蕴含识别任务最具挑战性的问题之一.

(3)受语义角色标注的影响,更容易进行句子成分的比较.如图5所示,我们的模型和基于BERT-wwm-ext的模型相比较,在中立关系的类别上预测错误的数量较多,这是受语义角色标注的影响,使得模型更容易进行句子成分的比较,具体示例如表6中3号句子对.但是上述情况出现的频率不高,对于含有复杂语义角色的句子,在推理关系时需要对句子成分进行充分比较.

5 总结

本文提出一种融合语义角色和自注意力机制的中文文本蕴含识别方法.首先把待识别语句分别输入到BERT-wwm-ext和SRL编码模块,然后在SRL-Attention模块中以自注意力机制的方法把编码后的SRL和BERT-wwm-ext的输出信息进行融合,最后对融合后的结果进行分类处理.实验结果表明该方法能够增强自注意力机制捕获句子语义的能力,并且提升模型在小规模数据集上的识别性能.在后续的研究工作中可以结合中文语法特点把融合方式细致化,进一步加强模型提取语义特征的能力.

参考文献

[1] 郭茂盛,张宇,刘挺.文本蕴含关系识别与知识获取研究进展及展望[J].计算机学报,2017,40(4):889-910.
GUO Mao-sheng, ZHANG Yu, LIU Ting. Research advances and prospect of recognizing textual entailment and knowledge acquisition[J]. Chinese Journal of Computers, 2017,40(4):889-910. (in Chinese)

[2] ANKUR P P, OSCAR T, DIPANJAN D, et al. A decomposable attention model for natural language inference [A]. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing [C]. Stroudsburg: Association for Computational Linguistics, 2016. 2249-2255.

[3] 袁里驰.利用配价信息的语义角色标注[J].电子学报,2017,45(10):2533-2539.
YUAN Li-chi. Semantic role labeling utilizing valence information[J]. Acta Electronica Sinica,2017,45(10):2533-2539. (in Chinese)

[4] SHI C, LIU S, REN S, et al. Knowledge-based semantic embedding for machine translation[A]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers)[C]. Berlin, Germany:Association for Computational Linguistics, 2016. 2245-2254.

[5] 杨震,范科峰,雷建军,等.基于语义的文本流形研究[J].电子学报,2009,37(3):557-561.
YANG Zhen, FAN Ke-feng, LEI Jia-jun, et al. Text manifold based on semantic analysis[J]. Acta Electronica Sinica,2009,37(3):557-561. (in Chinese)

[6] PETERS M E, NEUMANN M, IYYERM, et al. Deep contextualized word representations[A]. The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies[C]. Stroudsburg:Association for Computational Linguistics, 2018. 2227-2237.

[7] RANFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[OL]. https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_understanding_paper. pdf, 2019-10-20.

[8] DEVLIN J, CHANG M W, LEE K, et al. BERT:pre-training of deep bidirectional transformers for language understanding[A]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [C]. Stroudsburg: Association for Computational Linguistics, 2019. 4171-4186.

[9] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. arXiv Preprint, 2019, arXiv:1906. 08101.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[A]. Procceeding of 31st Conference on Neural Information Processing Systems (NIPS 2017) [C]. Long Beach, CA, USA:ACM, 2017. 6000-6010.

[11] BOWMAN S R, ANGELI G, POTTS C, et al. A large annotated corpus for learning natural language inference [A]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing [C]. Stroudsburg:Association for Computational Linguistics, 2015. 632-642.

[12] 翟延冬,王康平,张东娜,等.一种基于WordNet的短文本语义相似性算法[J].电子学报,2012,40(3):617-620.
ZHAI Yan-dong, WANG Kang-ping, ZHANG Dong-na, et al. An algorithm for semantic similarity of short text based on WordNet[J]. Acta Electronica Sinica, 2012, 40(3): 617-620. (in Chinese)

[13] 袁里驰.基于配价结构和语义依存关系的句法分析统计模型[J].电子学报,2013,41(10):2029-2034.
YUAN Li-chi. A statistical parsing model based on valence structure and semantic dependency[J]. Acta Electronica Sinica, 2013, 41(10):2029-2034. (in Chinese)

[14] 刘茂福,李妍,姬东鸿.基于事件语义特征的中文文本蕴含识别[J].中文信息学报,2013,27(5):129-136.
LIU Mao-fu, LI Yan, JI Dong-hong, et al. Event semantic feature based Chinese textual entailment recognition[J]. Journal of Chinese Information Processing, 2013, 27(5):

129 – 136. (in Chinese)
[15] 谭咏梅,刘姝雯,吕学强. 基于 CNN 与双向 LSTM 的中文文本蕴含识别方法[J]. 中文信息学报,2018,32(7):11 – 19.
TAN Yong-mei, LIU Shu-wen, LU Xue-qiang. CNN and BiLSTM based Chinese textual entailment recognition [J]. Journal of Chinese Information Processing, 2018, 32(7): 11 – 19. (in Chinese)
[16] ROCKTASCHEL T, GREFENSTETTE E, HERMANN K M, et al. Reasoning about entailment with neural attention [J]. arXiv Preprint, 2015, arXiv: 1509. 06664.
[17] WANG S, JIANG J. Learning natural language inference with LSTM [A]. The 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies [C]. Stroudsburg: Association for Computational Linguistics, 2016. 1442 – 1451.
[18] CHEN Q, ZHU X, LING Z, et al. Enhanced LSTM for natural language inference [A]. Proceedings of the 55th Annual Meeting of the Association for Computational [C]. Stroudsburg: Association for Computational Linguistics, 2017. 1657 – 1668.
[19] 刘广灿,曹宇,许家铭,等. 基于对抗正则化的自然语言推理[J]. 自动化学报,2019,45(8):1455 – 1463.
LIU Guang-can, CAO Yu, XU Jia-ming, et al. Natural language inference based on adversarial regularization [J]. Acta Automatica Sinica, 2019, 45(8): 1455 – 1463. (in Chinese)
[20] PAN B, YANG Y, ZHAO Z, et al. Discourse marker augmented network with reinforcement learning for natural language inference [A]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) [C]. Melbourne: Association for Computational Linguistics, 2018. 989 – 999.
[21] ZHANG Z, WU Y, LI Z, et al. I know what you want: Semantic learning for text comprehension [J]. arXiv Preprint, 2018, arXiv: 1809. 02794.
[22] BA J L, KIROS J R, HINTON G E. Layer normalization [J]. arXiv Preprint, 2016, arXiv: 1607. 06450.
[23] Che W, Li Z, Liu T. Ltp: A chinese language technology platform [A]. Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations [C]. Association for Computational Linguistics, 2010. 13 – 16.
[24] 中国计算语言学大会(CCL 2018)技术评测任务发布[R/OL]. http://www. cips-cl. org/static/CCL2018/call-evaluation. html, 2019-9-23.

作者简介

张志昌　男,1976 年 4 月出生,甘肃天水人. 教授、硕士生导师. 1998 年、2003 年和 2010 年分别在西北师范大学、西北工业大学、哈尔滨工业大学获工学学士、工学硕士和工学博士学位. 研究方向为自然语言处理,主要进行问答技术、医疗文本处理技术研究.
E-mail: zzc@ nwnu. edu. cn

曾扬扬　男,1996 年 4 月出生,河南项城人. 西北师范大学计算机科学与工程学院硕士研究生. 研究方向为自然语言处理.
E-mail: zeng2y@ 126. com