结合依存图卷积的中文文本相似度计算研究

胡书林[1]，张华军[1*]，邓小涛[2]，王征华[2]
（1.武汉理工大学自动化学院，湖北 武汉 430070；2.武汉大晟极科技有限公司，湖北 武汉 430075）

摘　要：目前中文文本相似度计算能够通过词嵌入技术在语义层面判别文本相似度，但通常忽略了文本中蕴含的丰富的句法结构信息，而以词为单位的中文句法分析与动态词嵌入模型中以字为单位的分词粒度不一致，使得当前大多数结合句法分析的研究只能使用静态词嵌入来表征词的语义向量。针对此问题，根据依存句法分析构建依存图，通过分词掩码映射与注意力混合池化的方法实现动态词嵌入表征词节点的语义特征，然后使用图卷积网络提取依存图中词节点之间的依存关系信息，最终读出依存图，将其作为句子的特征向量，从语义与句法 2 个层面计算句子间的相似度。在表示型与交互型 2 种结构模型上应用所提方法，并在 BQ_Corpus 与 ATEC 数据集上进行实验，结果显示，该模型的准确率最高分别达到 87.12%与 88.33%，结合依存句法信息后模型的各项评估指标均有提升。

关键词：图卷积神经网络；依存句法分析；动态词嵌入；文本相似度；注意力机制
中图分类号：TP18　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0068832

Similarity Calculation for Chinese Text Based on Dependency Graph Convolution

HU Shulin[1], ZHANG Huajun[1*], DENG Xiaotao[2], WANG Zhenghua[2]
(1. School of Automation, Wuhan University of Technology, Wuhan 430070, Hubei, China;
2. Wuhan DaSoundGen Technologies Co., Ltd., Wuhan 430075, Hubei, China)

【Abstract】 In the current landscape of Chinese text similarity computation, the use of word-embedding techniques enables discrimination of text similarity at the semantic level. However, this approach often overlooks the rich syntactic structural information inherent in texts. Chinese syntactic analysis at the word level is inconsistent with the granularity of the dynamic word-embedding models that operate at the character level. Consequently, most studies that combine syntactic analysis employ only static word embeddings to represent the semantic vectors of words. To address this issue, this study constructs a dependency graph based on syntactic dependency analysis. It employs a method involving tokenization mask mapping and attention-mix pooling to embed the semantic features of word nodes dynamically. Subsequently, a graph convolutional network is employed to extract the dependency relationship information among the word nodes in the dependency graph. The resulting dependency graph is then utilized as a feature vector for the sentence. The similarity between sentences is calculated from both semantic and syntactic perspectives. The proposed approach is applied to two model structures based on representation and interaction. Experimental evaluations are conducted using the BQ_Corpus and ATEC datasets. The experimental results demonstrate that the model achieves the highest accuracies of 87.12% and 88.33%, respectively. The incorporation of syntactic information leads to improvements in various model performance metrics.

【Key words】 graph convolution neural network; dependency syntactic parsing; dynamic word embedding; text similarity; attention mechanism

0　引言

文本相似度计算是在语义层面上计算 2 个句子是否相似，是自然语言处理中的一个核心任务，其研究可以应用到信息检索、自动问答、机器翻译、对话系统、复述问题等其他自然语言处理任务中[1]。目前主要的文本相似度计算方法是在语义层面通过词嵌入技术将文本解析为低维稠密向量，并计算向量之间的余弦距离或欧氏距离得到文本之间的相似度得分[2]。词嵌入分为静态词嵌入与动态词嵌入 2 种方法[3]，这 2 种方法都能将文本映射为低维稠密向量表示，但动态词嵌入通过利用大规模文本数据的上下文信息来丰富词汇表示，解决了静态词嵌入无法处理一词多义的问题。文献[4]最早尝试了动态词嵌入的词向量模型 CoVe，通过训练一个长短期记忆（LSTM）编码器

收稿日期：2023-11-13　　**修回日期：**2024-01-30
基金项目：湖北省重点研发计划项目（2022BAA051）。
通信作者 E-mail：* zhanghj@whut.edu.cn

计算机工程 www.ecice06.com

生成根据上下文变化的词嵌入，再应用到下游任务中；文献[5]提出了 ELMo 语言模型嵌入方法，使模型能在大量未标注语料上进行训练，并构建一个深度双向语言模型来提取词向量表示；文献[6]提出了基于 Transformer 的预训练模型 BERT，模型在单词掩码与下一句预测这 2 种新的无监督任务上进行预训练，使模型更准确地捕获了丰富的上下文语义信息。目前，效果较好的动态词嵌入模型包括但不限于预训练语言模型 RoBERT[7]、GPT[8]等。

虽然词嵌入技术能够提取文本的深层语义信息，但通常忽略了自然语言中拥有的丰富语法和语义结构信息。句法结构作为句子的重要组成部分，包含了词之间丰富的语法关系，能够辅助解析文本的语义信息。常用的句法分析包含短语结构分析与依存句法分析，短语结构分析关注句子如何生成，而依存句法分析更关注句子中词的依存关系。这类信息都属于非欧氏空间数据，其邻居节点数量不同，不具有平移不变性，因此适用于图数据结构进行表示，能够通过图神经网络进行学习训练。

目前，在自然语言处理中已有许多依存图结合图神经网络的研究[9]，但由于中文以词为单位的句法分析规则与词嵌入以字符为单位的分词编码规则不一致，导致图神经网络中词节点的语义特征难以表示，目前大多数相关研究都使用静态词嵌入技术，或者自行构建词向量表示，例如：文献[10]使用 Word2Vec[11]静态词嵌入生成语义信息的词向量，结合依存分析与图神经网络完成情感分析任务；文献[12]通过 gazetteers 和 bigram 表示的词嵌入向量融合来初始化向量，然后结合依存分析与自适应门控图神经网络实现中文命名实体识别；文献[13]使用百度百科语料库训练语言模型得到词向量表，根据词向量表得到分词的向量表示，再结合图卷积网络完成实体关系的联合抽取。

目前，对于词粒度的语义表达已有相关研究，例如文献[14]提出基于词对齐注意力的方法，实现中文词的语义表征，该方法能够减轻当前预训练模型基于字符级分割的错误信息传播，实现词语义的准确表达。

本文使用分词掩码映射的方法实现图神经网络中词节点的语义表征，充分发挥动态词嵌入模型在语义表征中的效果，在文本语义解析中融合文本的句法结构信息，使模型能够在语义与句法结构 2 个层面计算文本间的相似度，提高模型评估的准确性。通过在开源数据集 BQ_Corpus 与 ATEC 上进行实验，验证结合句法结构信息后文本相似度计算的准确率是否得到提升。

1 相关工作

图神经网络是一种专门处理图数据的深度学习模型，能够捕捉节点之间的复杂关系并对信息进行聚合传递，因此，图神经网络在自然语言处理中拥有广泛应用。例如，图神经网络可以解决跨句的多元关系语义提取问题，在推荐系统方面可以有效地完成矩阵补全问题[9]。目前，已有相关研究使用图神经网络对句子依存关系进行学习来提高模型的解析能力，例如文献[15]使用图卷积网络将句法与语义信息融合于词嵌入学习中，文献[16]将多个依存分析结果集成到一个图中进行学习，以减轻依存解析错误的影响，实现更准确的情感分析。上述研究成果表明，图神经网络能够有效捕捉句子内各个组成部分之间的依存结构、语法规则和语义信息等复杂关系，将依存分析等句法结构与图神经网络相结合后，可进一步提高模型在文本分析和处理任务中的性能。

传统的文本相似度计算方法是在词汇层面通过人工设计特征与浅层模型的组合来完成任务，例如 TF-IDF(Term Frequency-Inverse Document Frequency)算法[17]根据词频与文档长度等特征计算文本的特征向量及其之间的相似度。但这些方法都忽略了词的序列关系与上下文信息，无法有效地捕捉语义信息。而基于深度学习的文本相似度计算方法能够在语义层面学习文本的蕴含信息，CNN、RNN 等神经网络结构能够进一步提取文本上下文信息，更深层次地解析文本的整体语义。而随着预训练语言模型的快速发展，例如 BERT 等基于 Transformer 架构的语言模型在大量文本语料上进行无监督训练，使得模型的自然语言解析能力取得了显著的提升，且能够在其他文本数据上进行迁移学习，从而更准确地获取文本的语义信息。

目前，文本相似度匹配主要有交互型与表示型 2 种模型结构[18]，如图 1 所示。其中，BERT 可替换为其他词嵌入模型。交互型是指将 2 个句子拼接成一段文本并进行解析分类，由于文本从最开始就获得了充分的交互比较，例如文献[19]将文本匹配视为图像识别问题，通过对句子对的 embedding 进行匹配交互，然后用二维卷积提取特征，获取了较好的语义解析效果。表示型是指将 2 个句子分别解析为语义向量，再进行交互以计算相似度，例如文献[20]

计算机工程 www.ecice06.com

提出的 DSSM (Deep Structured Semantic Models) 双塔模型,将文本对分别解析,并在语义空间上映射为特征向量表示,最后计算两者的相关性分数。这类方法能够提前计算并缓存句子的语义向量,只需要在最后进行融合,因此准确率虽不如交互式,但计算效率较高。

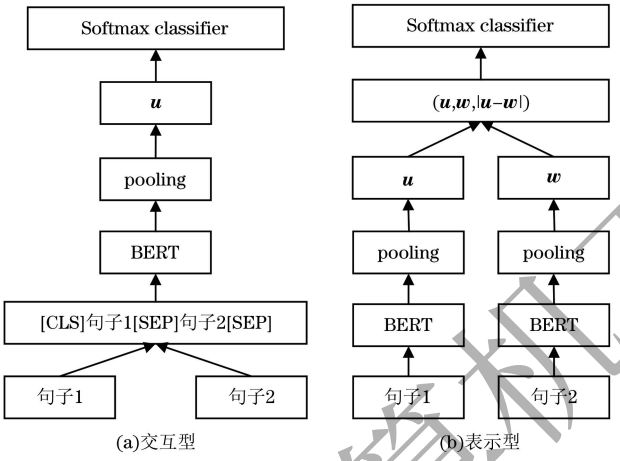


图 1 文本匹配的 2 种模型结构

Fig. 1 Two model structures for text matching

随着深度学习的发展,预训练模型学习了大规模文本数据的语言表示,在下游任务中进行微调便可取得较好的识别效果。如文献[21]构建 Sentence-BERT 孪生网络模型结构,用于计算句子的相关程度,这种实现方法以及预训练模型的广泛使用使得文本相似度计算变得愈发快速且准确。

基于上述研究,本文结合预训练模型的动态表征,通过依存图卷积网络来学习词之间的依存关系,进而提高文本相似度计算效果,并在交互型与表示型 2 种结构中分别应用本文方法进行实验以验证模型的准确率。

2 模型结构

本文提出的依存图卷积模型主要包含图表示层、图卷积层与图读出层,模型结构如图 2 所示。输入的句子经过预训练模型分词编码与依存分析后,构建依存图表示,然后将图传入图卷积层学习句子的语义与依存关系,卷积网络提取的图特征作为句子的特征向量,参与后续的相似度分类层计算。

2.1 图表示层

2.1.1 分词掩码映射

本文以 BERT 为例,使用预训练语言模型对句子分词编码。当前大部分的中文预训练模型都是以字为基本单位,如 BERT 模型对包含 $N \in \mathbb{R}$ 个字符的句子进行分词,并在句首和句尾分别添加特殊

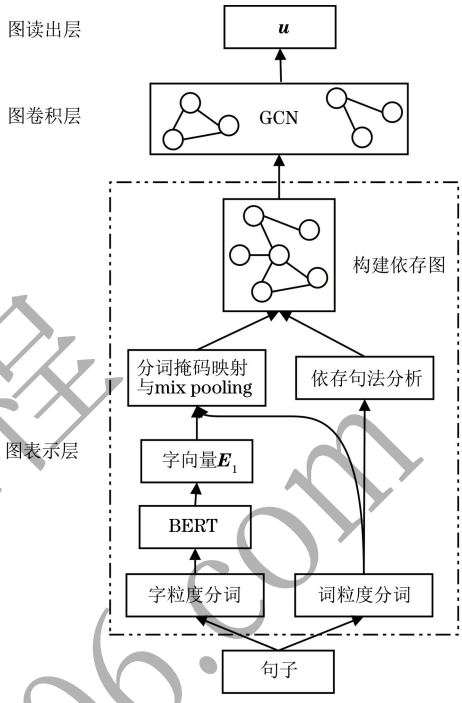


图 2 依存图卷积模型结构

Fig. 2 Dependency graph convolutional model structure

符号[CLS]与[SEP]得到以字为粒度的分词结果 tokens1:

$$\text{tokens1} = [[\text{CLS}], \text{word}_1, \text{word}_2, \cdots, \text{word}_N, [\text{SEP}]] \quad (1)$$

式中:[CLS]代表整个文本的语义表示;[SEP]为句子之间的分隔符。

tokens1 被送入 BERT 模型编码并学习每个单词的上下文表达,最后将 BERT 模型最后一层的输出 $\boldsymbol{E} \in \mathbb{R}^{(N+2)\times d}$ 作为输入句子的语义特征表示:

$$\boldsymbol{E} = [\boldsymbol{e}_{\text{CLS}}, \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_N, \boldsymbol{e}_{\text{SEP}}]^{\text{T}} \quad (2)$$

式中:$\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\cdots$, $\boldsymbol{e}_N$ 分别对应分词结果 word_1, $\text{word}_2, \cdots, \text{word}_N$ 中 N 个字符的语义向量表示,语义向量的维度为 d 。

由于中文预训练语言模型以字为基本单位进行分词编码,而依存分析以词为单位进行分析,分词结果存在字词混合的情况,因此需要针对多粒度的词进行进一步处理来获取词的语义向量表示,最终作为图卷积层的节点特征输入。

本文使用哈尔滨工业大学开发的语言技术平台(LTP)[22]进行中文分词与依存句法分析,得到分词结果 tokens2:

$$\text{tokens2} = [\text{words}_1, \text{words}_2, \cdots, \text{words}_M] \quad (3)$$

式中:$M \in \mathbb{R}$ 为 tokens2 的分词长度。

本文通过构建分词掩码特征矩阵 $\boldsymbol{M}_{\text{Mask}} \in \mathbb{R}^{M\times(N+2)}$ 将预训练模型分词结果 tokens1 与 LTP 中文分词结果 tokens2 相互映射,为后续词级的语

计算机工程 www.ecice06.com

义表示计算提供依据。如图 3 所示，根据字粒度分词结果找到每个字在词粒度分词结果中的位置，并将相应的元素设置为 1，最终得到分词掩码 $\boldsymbol{M}_{\text{Mask}}$ 。

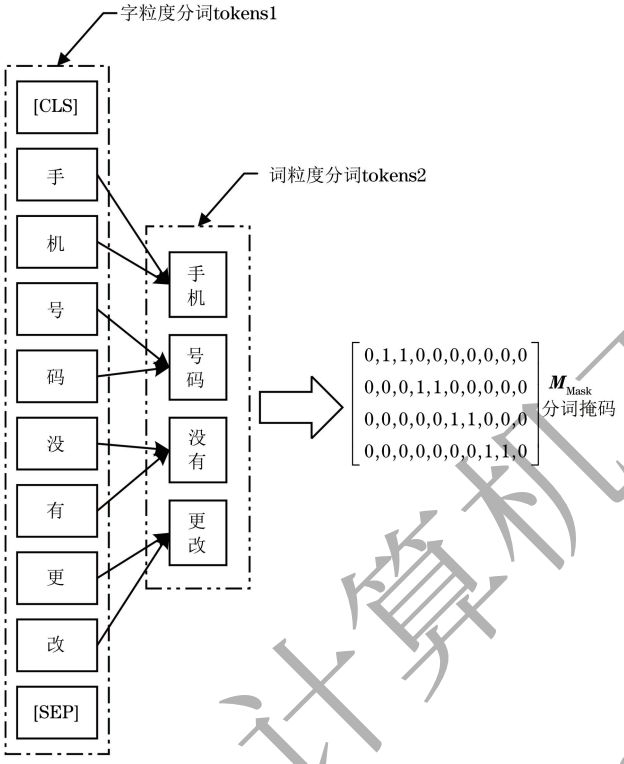


图 3 分词掩码映射示例

Fig. 3 Segmentation mask mapping example

2.1.2 注意力权重混合池化

由于依存图是以词为节点、以词之间的依存关系为边进行构建的，因此需要得到词的语义表示信息作为节点特征输入。本文结合 2.1.1 节中的分词掩码，将字符级语义向量 $\boldsymbol{E}$ 根据自注意力权重求和并池化输出。该方法根据上下文中每个字的重要性来动态地组合字的语义信息，以生成整个词的语义表示，能提高中文词语义表示的准确性，减轻语义偏差问题，并作为后续依存图卷积层计算的节点特征输入。

自注意力机制允许模型在输入序列内部不同位置的元素之间建立权重关系，以便更好地捕捉序列内的依赖关系和上下文信息[23]。对于输入的句子序列向量 $\boldsymbol{E}\in\mathbb{R}^{(N+2)\times d}$，有：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{E}\boldsymbol{W}^{\boldsymbol{Q}}\\ \boldsymbol{K}=\boldsymbol{E}\boldsymbol{W}^{\boldsymbol{K}}\end{cases}\tag{4}$$

式中：$\boldsymbol{Q}\in\mathbb{R}^{(N+2)\times d}$、$\boldsymbol{K}\in\mathbb{R}^{(N+2)\times d}$ 分别为计算注意力的查询与键值矩阵；$\boldsymbol{W}^{\boldsymbol{Q}}\in\mathbb{R}^{d\times d}$、$\boldsymbol{W}^{\boldsymbol{K}}\in\mathbb{R}^{d\times d}$ 为可训练的参数矩阵。所有字符级的语义向量表示都经过自注意力操作，得到字符级的自注意力得分输出 $\boldsymbol{A}_{\text{score}}\in\mathbb{R}^{(N+2)\times(N+2)}$：

$$\boldsymbol{A}_{\text{score}}=\text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}\right)\tag{5}$$

为了方便解释，本文将自注意力得分计算结果 $\boldsymbol{A}_{\text{score}}$ 与字符级语义向量 $\boldsymbol{E}$ 对齐：

$$\boldsymbol{A}_{\text{score}}=[\boldsymbol{a}_{\text{CLS}},\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_N,\boldsymbol{a}_{\text{SEP}}]\tag{6}$$

式中：$\boldsymbol{a}_i\in\mathbb{R}^N$ 表示第 i 个字符的自注意力得分向量。随后将自注意力得分作为权重，与字符级语义向量进行点积计算，并将计算结果池化输出作为词级语义向量表示输出。图 3 例句中“手机”的语义向量为 $\boldsymbol{e}'_1$，而“手”、“机”的语义向量分别为 $\boldsymbol{e}_1$、$\boldsymbol{e}_2$，则“手机”的词向量应为：

$$\boldsymbol{e}'_1=P(\boldsymbol{a}_1\boldsymbol{e}_1+\boldsymbol{a}_2\boldsymbol{e}_2)\tag{7}$$

式中：P 表示池化函数。本文使用混合池化方法计算 tokens2 中词对应的语义向量 $\boldsymbol{E}'\in\mathbb{R}^{M\times d_{\text{dim}}}$：

$$\boldsymbol{e}'_k=\lambda\cdot\text{Maxpooling}(\boldsymbol{M}_{\text{Mask}}[k]\boldsymbol{A}_{\text{score}}\boldsymbol{E})+(1-\lambda)\cdot\text{Meanpooling}(\boldsymbol{M}_{\text{Mask}}[k]\boldsymbol{A}_{\text{score}}\boldsymbol{E})\tag{8}$$

$$\boldsymbol{E}'=[\boldsymbol{e}'_1,\boldsymbol{e}'_2,\cdots,\boldsymbol{e}'_k,\cdots,\boldsymbol{e}'_M]^{\text{T}}\tag{9}$$

式中：$k\in[1:M]$；$\lambda\in\mathbb{R}$，λ 是加权可训练变量，用于平衡最大池化和平均池化的计算结果。最终，$\boldsymbol{E}'$ 作为词节点特征输入，用于后续图卷积层的学习计算。

2.1.3 依存图构建

LTP 在对句子进行依存分析后，根据依存分析结果，将词作为节点、将依存关系作为边构建依存图，并对每个节点添加自环边，最终得到该句子的依存图表示 $G=(V,E)$。V 代表图中节点的集合，E 代表图中边的集合，由于边包含多种类型的关系且为有向边，因此依存图 G 为异构有向图。例句“手机号码没有更改”的依存分析结果如图 4 所示，每个

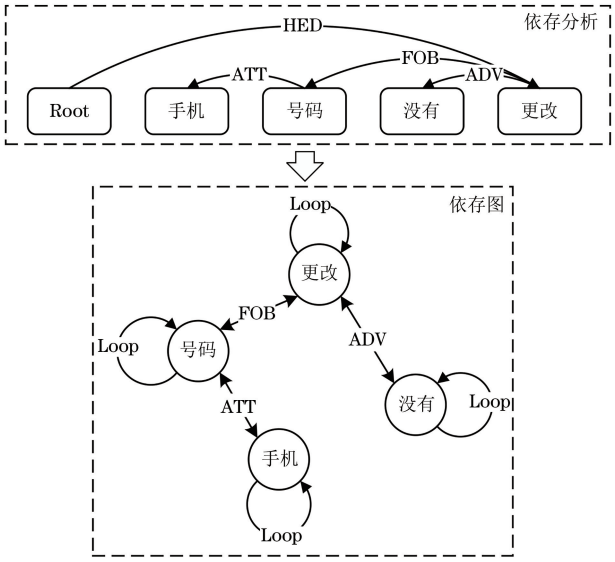


图 4 依存图构建示例

Fig. 4 Dependency graph construction example

计算机工程 www.ecice06.com

词作为依存图中的节点，通过依存关系与其他词节点相连。为了使图特征表现得更加明显，在原始的边关系中增加依存关系的反方向边与自环边，最终完成依存分析到依存图的构建。在图 4 中，依存关系边及其反向边表示为一个双向的关系边，自环边表示为一个名为“Loop”的关系边。

2.2 图卷积层

图卷积层将根据句子依存分析构建的依存图 G 计算每个词节点的邻接矩阵，并依照词节点的语义向量表达 $\boldsymbol{E}'$，对每个词节点的语义向量与其邻接节点的词向量进行卷积操作，得到新的词向量表示，以此来学习词的语义信息以及词之间的依存关系。图卷积神经网络允许信息随着图中的边不断向远处传播，通过增加图卷积的层数可以帮助模型解决距离较远词的信息传递问题。但由于不同关系边需要计算不同的邻接矩阵，且依存图 G 为异构图，因此在图卷积运算中依存图 G 被拆分为一系列子图，每种子图对应一种关系，并在对应关系的图卷积网络上进行运算以更新信息：

$$\boldsymbol{h}_{\mathrm{dst}}^{(l+1)}=\underset{r\in R,r_{\mathrm{dst}}=\mathrm{dst}}{\mathrm{AGG}}\left(f_r\left(g_r,\boldsymbol{h}_{r_{\mathrm{src}}}^{l},\boldsymbol{h}_{r_{\mathrm{dst}}}^{l}\right)\right) \tag{10}$$

式中：AGG 是聚合函数；f_r 是对应每个关系 r 的图卷积网络模块；g_r 是图 G 的关系 r 对应的子图；$\boldsymbol{h}_{r_{\mathrm{src}}}^{l}$ 是图卷积层中第 l 层关系 r 的头节点；$\boldsymbol{h}_{r_{\mathrm{dst}}}^{l}$ 是图卷积层中第 l 层关系 r 的尾节点。

2.3 图读出层

依存图 G 在图卷积网络更新信息后整图读出，图的读出公式如下：

$$\boldsymbol{h}_g=\frac{1}{|V|}\sum_{v\in V}\boldsymbol{h}_v \tag{11}$$

式中：$\boldsymbol{h}_g$ 是图 G 的特征表示；V 是图 G 中节点的集合；$\boldsymbol{h}_v$ 是节点 v 的特征。所有节点特征被聚合并取均值后作为句子的语义向量表示，参与后续相似度计算。

交互型与表示型模型结构中图读出后的计算存在不同之处，在表示型的模型结构中 2 个句子的图分别读出，假设得到的 2 个依存图读出后的特征表示分别为 $\boldsymbol{u}$ 和 $\boldsymbol{w}$，本文将 $\boldsymbol{u}$、$\boldsymbol{w}$ 以及 2 个特征差的绝对值 $|\boldsymbol{u}-\boldsymbol{w}|$ 拼接后，由 Softmax 函数来评估 2 个句子之间的相似度概率：

$$y=\mathrm{Softmax}(\mathrm{Concat}(\boldsymbol{u},\boldsymbol{w},|\boldsymbol{u}-\boldsymbol{w}|)) \tag{12}$$

基于交互型的依存图卷积在完成图卷积层的信息更新后，将 2 个依存图合并读出得到整图特征表示 $\boldsymbol{u}$，然后由 Softmax 函数直接评估 2 个句子之间的相似度概率：

$$y=\mathrm{Softmax}(\boldsymbol{u}) \tag{13}$$

3 实验验证

3.1 数据集

本文使用的实验数据集为 BQ_Corpus 数据集与 ATEC 数据集。BQ_Corpus 为银行金融领域的问题匹配数据集，包括从一年的线上银行系统日志中抽取的问题 pair 对，BQ_Corpus 的训练集样本量为 10 万条句子对，验证集为 1 万条，测试集为 1 万条；ATEC 为蚂蚁金服 NLP 挑战数据集，训练集样本量为 6 万多条，验证集为 2 万条，测试集为 2 万条。由于需要在交互型与表示型 2 种模型上进行实验，交互型模型输入的句子需要拼接处理，因此预训练模型在分词时最大序列长度需要分别统计分析。

2 个数据集分词后的最大序列长度如图 5 所示。train_single 表示训练集中单句分词后的最大序列长度，train_concat 表示训练集中句子对拼接分词后的最大序列长度。由图 5 可知，只有 BQ_Corpus 的训练集中部分句子分词后序列长度超过 128，其余数据集单句序列长度均小于等于 128，所有数据集的句子序列长度均小于 256。因此，在表示型结构的模型实验中，最大序列长度设置为 128，舍弃数据集中超出最大序列的句子对；在交互型结构的模型实验中，最大序列长度设置为 256。

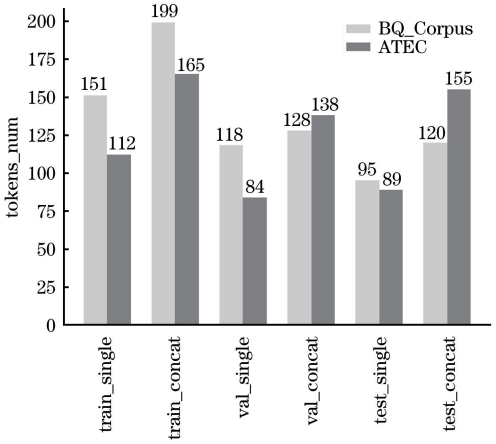


图 5 最大序列长度统计

Fig. 5 Maximum sequence length statistics

3.2 参数设置

本文实验基于 PyTorch 深度学习框架实现，使用深度图谱库（DGL）实现图数据的加载与图卷积层的构建，模型具体参数设置见表 1。模型最终的评估结果为 0 与 1 的概率，若分类结果为 1 的概率更大，则代表 2 个句子相似，反之则不相似。

计算机工程 www.ccieeo.com

表 1 实验参数设置

Table 1 Experimental parameter settings

参数	数值
词嵌入维度	768
表示型 max_len	128
交互型 max_len	256
学习率	1×10^{-5}
Dropout	0.1
GCN 层数	3
Batch size	64
Epoch	5

3.3 评估指标

本文使用准确率 accuracy 和 F1 值这 2 个评估指标对相似度计算结果进行评价，其中准确率与 F1 值的计算方法如式(14)～式(17)所示：

$$V_{\text{accuracy}}=\frac{N_{\text{TP}}+N_{\text{TN}}}{N_{\text{TP}}+N_{\text{TN}}+N_{\text{FP}}+N_{\text{FN}}} \tag{14}$$

$$V_{\text{precision}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}} \tag{15}$$

$$V_{\text{recall}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}} \tag{16}$$

$$F_1=\frac{2\cdot V_{\text{precision}}\cdot V_{\text{recall}}}{V_{\text{precision}}+V_{\text{recall}}} \tag{17}$$

式中：N_{TP} 表示将相似的句子对正确预测为相似的样本数量；N_{TN} 表示将不相似的句子对成功预测为不相似的样本数量；N_{FP} 表示将不相似的句子对错误预测为相似的样本数量；N_{FN} 表示将相似的句子对错误预测为不相似的样本数量。accuracy 表示预测正确的句子对占样本总数的比例，F1 值表示精确率与召回率的调和平均值。

3.4 对比实验

本文提出的依存图卷积模型是在交互型结构与表示型结构上使用动态词嵌入融合依存句法信息而进行的改进。表示型依存图卷积网络使用 Siamese-GCN 表示，交互型依存图卷积网络使用 Cross-GCN 表示。为了验证本文方法的有效性，本文设置了以下基线模型进行实验：

1）Word2Vec-Baseline。使用预训练的 Word2Vec 生成词向量表示，均值池化后得到句子的语义向量，并进行相似度计算。

2）BiMPM[24]。采用双向多角度匹配，基于交互型结构对句子对之间的单元进行相似度计算，解决之前交互模型匹配不充分的问题。

3）ESIM[25]。增强了序列推断能力的文本匹配模型，综合应用了 BiLSTM 和注意力机制，精细地设计序列式的推断结构，更全面地实现了局部推理与全局推理。

4）RE2[26]。通过保持序列间对齐的原始点对齐特性、先前对齐特性和上下文特性等 3 个特征，实现快速高效的通用文本匹配。

5）Siamese-BERT[21]。Siamese-BERT 也称 Sentence-BERT 或 Bi-Encoder，通过构建孪生编码器网络实现 2 个编码器共享权重，最终在输出层比较 2 个句子的语义向量相似度。

6）Cross-BERT。交互型网络结构，与表示型的 Siamese-BERT 相对应，对句子对拼接后编码解析得到的语义向量进行计算得到相似度结果。

7）BERT-whitening[27]。采用传统机器学习中的白化操作直接校正句向量的协方差矩阵，增强句子表示的各向同性，以此提升文本相似度计算的效果，降低存储成本，加快模型检索速度。

8）RoFormer[28]。RoFormer 中设计了一种名为旋转位置嵌入（RoPE）的新方法，用旋转矩阵对绝对位置进行编码，同时在自注意力公式中加入了显式相对位置依赖性，以提高 Transformer 架构的性能。

ERNIE[29-30]是基于飞桨开源的持续学习的语义理解框架，该预训练模型在大型数据集合中进行多任务增量训练，学习了训练数据中的词法信息、语法信息与语义信息。为了进一步探究融合句法信息对文本相似度计算效果的影响，本文将实验中的动态词嵌入模型替换为 ERNIE 进行对比。

3.5 结果分析

3.5.1 对比实验结果分析

对比实验结果如表 2 所示，从中可以看出，本文提出的方法在 BQ_Corpus 与 ATEC 数据集上的准确率分别达到了 87.12%、88.33%，F1 值分别达到了 86.80%、87.42%，与使用静态词嵌入的 Word2Vec-Baseline 相比，本文方法各项评估指标均显著提高，这验证了动态词嵌入对句子语义的表征效果明显优于传统的静态词嵌入方法。而与 BiMPM、ESIM、RE2 等自行构建编码层的相似度计算方法相比，本文方法各项评估指标也有不同程度的提升，这一结果也表明，比起自主构建的编码层，基于预训练模型的编码器能够准确地解析句子中丰富的语义信息。

Siamese-GCN 与 Cross-GCN 通过分词掩码映射的方法表示中文词的语义向量，并通过图卷积网络学习句子的依存句法信息。在加入依存句法特征信息后，表示型结构的 Siamese-GCN（BERT）相较 Siamese-BERT 在 2 个数据集上的准确率分别提升了2.05、1.27百分点，Siamese-GCN（ERNIE）的

计算机工程 www.ecice06.com

表 2 对比实验结果

Table 2 Comparative experimental results %

模型	BQ_Corpus		ATEC	
	accuracy	F1 值	accuracy	F1 值
Word2Vec-Baseline	73.28	72.23	74.89	73.60
BiMPM[24]	79.69	79.30	81.58	81.09
ESIM[25]	80.04	79.93	81.88	81.35
RE2[26]	80.09	79.81	82.04	81.67
Siamese-BERT[21]	82.56	82.03	85.07	84.88
Siamese-ERNIE	83.63	83.07	85.89	85.22
Cross-BERT	83.66	83.11	85.68	85.00
Cross-ERNIE	84.97	84.11	86.54	86.12
BERT-whitening[27]	84.02	83.49	85.96	85.17
RoFormer[28]	85.14	84.71	86.77	86.23
Siamese-GCN (BERT)	84.61	84.07	86.34	85.93
Siamese-GCN (ERNIE)	85.34	84.86	86.98	85.99
Cross-GCN (BERT)	86.46	86.18	87.62	87.04
Cross-GCN (ERNIE)	87.12	86.80	88.33	87.42

准确率相较 Siamese-ERNIE 分别提升了 1.71、1.09 百分点；交互型结构的 Cross-GCN（BERT）相较 Cross-BERT 在 2 个数据集上的准确率分别提升了 2.8、1.94 百分点，Cross-GCN（ERNIE）相较 Cross-ERNIE 在 2 个数据集上的准确率分别提升了 2.15、1.79 百分点。上述实验结果表明，在保持模型结构一致的前提下，引入依存句法信息可以使模型从语义和句法 2 个层面来学习句子的特征表示，并抽取更多的句子间特征来计算文本相似度。因此，本文模型的评估效果优于仅考虑语义层面的模型，这显示出依存句法信息在文本相似度计算中的积极意义。

BERT-whitening 在 2 个数据集上的准确率相比 Cross-BERT 模型分别提升 0.36、0.28 百分点，这说明了白化操作确实能增强句子表示的各向同性，在一定程度上提升相似度计算的效果。但 BERT-whitening 在 2 个数据集上的准确率相比 Cross-GCN (BERT)模型分别降低 2.44、1.66 百分点，分析其原因可能是学习了句法信息后的语义表示进一步增强了句子间的各向同性，使相似度计算性能获得了增益。RoFormer 相较 Cross-BERT 模型在 2 个数据集上的准确率分别提高了 1.48 与 1.09 百分点，但相较 Cross-GCN（BERT）模型分别降低 1.32、0.85 百分点，分析其原因可能是在交互式模型中的输入为句子对拼接后的长文本，RoFormer 使用旋转位置编码的方式来增强模型对较长文本的解析能力，这使得相似度计算准确率获得较大提高；而 Cross-GCN 采用图结构数据表示分词后的 tokens，这在一定程度上会增加数据处理的步骤，但更容易在句法层面捕获长文本中 tokens 之间的依赖关系，丰富句子的特征表示，使 Cross-GCN 表现出更好的性能。

预训练模型由 BERT 替换为 ERNIE 后，Siamese-GCN 与 Cross-GCN 均获得不同程度的性能提升，这表明 ERNIE 在多任务增量训练后进一步提升了模型语义解析能力。但在预训练模型替换后，依存图卷积层对文本相似度计算任务的性能提升程度有所下降，例如在 BQ_Corpus 数据集中，Cross-GCN (BERT) 相较 Cross-BERT 提升最大，准确率与 F1 值分别提高了 2.8、3.07 百分点，而 Cross-GCN (ERNIE)相较 Cross-ERNIE 在 BQ_Corpus 上准确率与 F1 值分别提高了 2.15、2.69 百分点。本文分析这是由于预训练模型的训练策略不同所导致的，ERNIE 参考多任务学习的框架，所有任务共享低层网络以学习更通用的文本特征，并同时训练多个任务来提取词法、语法和语义方面信息，这使得模型已具有解析部分句法结构的能力，因此削弱了图卷积层对依存句法特征的学习效果。除此以外，在引入依存图卷积层后，交互型结构模型获得的性能提升程度高于表示型结构模型，分析原因是交互式依存图卷积模型在图读出层获取句子对特征时又进行一次交互计算，这使得模型能够更深入地解析句子对之间的内在联系，进一步丰富相似度评估时所需的上下文信息，增强了模型的解析能力与精准度。

3.5.2 消融实验结果分析

为了验证注意力权重混合池化在词向量语义表示中的效果，本文设置消融实验，在 Cross-GCN (ERNIE)模型上采用不同池化方法进行对比验证，实

计算机工程 www.ecice06.com

验结果如表 3 所示。从中可以看出，最大池化、均值池化与混合池化带来的性能差距较小，各评估指标相差在 0.2 百分点以内，但总体上混合池化还是优于其他 2 种池化方法。注意力权重对模型性能带来了较大提升，3 种池化方法在加入注意力权重后各项评估指标均获得了明显提升，其中最大池化的提升幅度最大，在 2 个数据集上的准确率分别提高了 1.01、1.19 百分点。分析原因是词向量的语义表示并非是字向量特征的简单叠加，而注意力机制可以动态地对底层特征进行加权，强调信息传递中重要的部分，弱化不重要的部分，更有效地表示词的复杂语义特征，对提高模型的准确率和鲁棒性带来实质性的帮助。

表 3 消融实验结果

Table 3 Results of ablation experiment %

池化方法	BQ_Corpus		ATEC	
	accuracy	F1 值	accuracy	F1 值
最大池化	86.01	85.72	87.12	86.77
均值池化	86.13	85.78	87.13	86.83
混合池化	86.19	85.89	87.25	86.86
注意力权重最大池化	87.02	86.72	88.31	87.39
注意力权重均值池化	87.07	86.78	88.33	87.40
注意力权重混合池化	87.12	86.80	88.33	87.42

3.5.3 GCN 层数分析

为了进一步探究 GCN 层数对模型性能的影响，本文分别在 BQ_Corpus 与 ATEC 数据集上进行对比实验。如图 6(a)、图 6(b)所示，在 BQ_Corpus 数据集上，当 GCN 层数逐渐增加到 3 时，模型性能综合表现最优；当 GCN 层数超过 3 时，模型的各项性能指标开始下降。如图 7(a)、图 7(b)所示，在ATEC数据集上，除了交互型的Cross-GCN

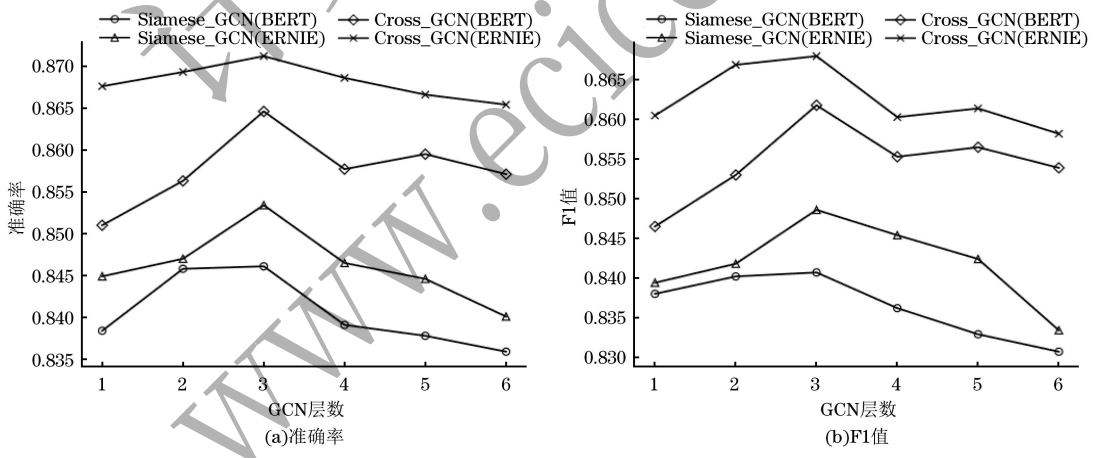


图 6 BQ_Corpus 数据集上 GCN 层数对模型性能的影响

Fig.6 The impact of GCN layers on model performance on the BQ_Corpus dataset

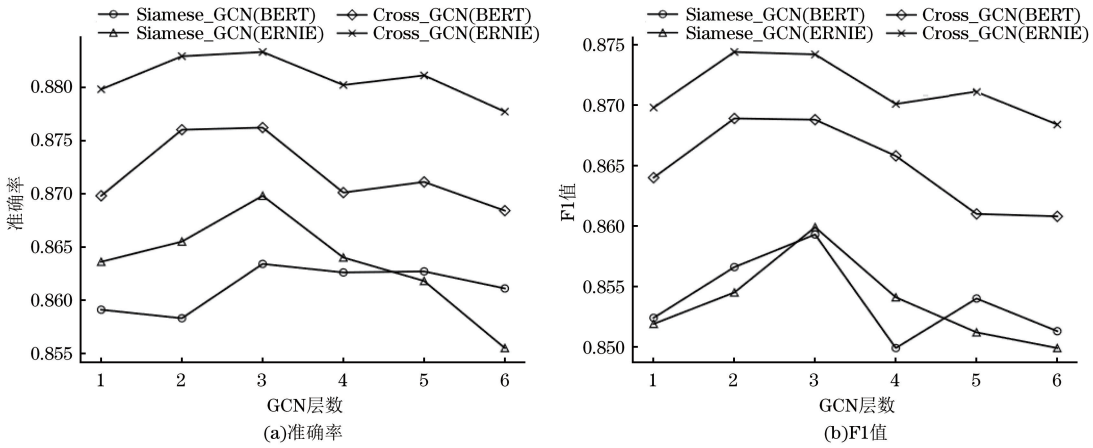


图 7 ATEC 数据集上 GCN 层数对模型性能的影响

Fig.7 The impact of GCN layers on model performance on the ATEC dataset

计算机工程 www.ecice06.com

模型在 GCN 层数为 2、3 时各项评估指标相差较小，其余模型在 GCN 层数增加到 3 时综合表现最优，层数超过 3 时，模型的各项性能指标开始回落。因此，本文将模型的 GCN 层数均设置为 3。

4 结束语

为了解决在依存图神经网络中难以使用动态词嵌入表征中文词向量的问题，本文提出了分词掩码结合自注意力权重混合池化的方法，实现依存图卷积网络的文本语义解析，将依存句法信息融入句子的语义表征中，以提高模型在文本相似度预测任务中的性能。首先，将动态词嵌入模型中以字符为粒度的分词结果与中文中以词为粒度的分词结果通过分词掩码实现映射；然后，针对词嵌入解析的序列向量，根据自注意力权重混合池化实现词的语义向量表征；接着，将词向量作为节点特征输入到依存图卷积网络中，以提取词之间的依存句法信息；最后，整图读出作为句子的特征表示，计算句子对的相似度。本文在表示型与交互型 2 种结构模型中应用所提方法，并在开源的 BQ_Corpus 与 ATEC 数据集上进行实验，测试结果验证了本文方法的有效性。

未来将在其他文本匹配任务中评估所提模型，例如释义识别与问答匹配任务。此外，将增大数据量并拆分数据集来进一步评估模型在不同领域数据集下的性能表现，探究模型解析不同长度句子的性能差异。

参考文献

[1] 庞亮，兰艳艳，徐君，等. 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，et al. A survey on deep text matching[J]. Chinese Journal of Computers，2017，40(4)：985-1003. (in Chinese)

[2] CHOWDHURY G G. Introduction to modern information retrieval[M]. [S. l.]：Facet Publishing，2010.

[3] WANG Y X，HOU Y T，CHE W X，et al. From static to dynamic word representations：a survey[J]. International Journal of Machine Learning and Cybernetics，2020，11(7)：1611-1630.

[4] MCCANN B，BRADBURY J，XIONG C，et al. Learned in translation：contextualized word vectors[EB/OL]. [2023-10-05]. https：//arxiv. org/abs/1708. 00107v2.

[5] PETERS M，NEUMANN M，IYYER M，et al. Deep contextualized word representations[EB/OL]. [2023-10-05]. https：//arxiv. org/pdf/1802. 05365.

[6] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/1810. 04805v2.

[7] LIU Y H，OTT M，GOYAL N，et al. RoBERTa：a robustly optimized BERT pretraining approach[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/1907. 11692v1.

[8] RADFORD A，NARASIMHAN K. Improving language understanding by generative pre-training[EB/OL]. [2023-10-05]. https：//gwern. net/doc/www/s3-us-west-2. amazonaws. com/d73fdc5ffa8627bce44dcda2fc012da638ffb158. pdf.

[9] 马帅，刘建伟，左信. 图神经网络综述[J]. 计算机研究与发展，2022，59(1)：47-80.
MA S，LIU J W，ZUO X. Survey on graph neural network[J]. Journal of Computer Research and Development，2022，59(1)：47-80. (in Chinese)

[10] 范涛，王昊，吴鹏. 基于图卷积神经网络和依存句法分析的网民负面情感分析研究[J]. 数据分析与知识发现，2021，5(9)：97-106.
FAN T，WANG H，WU P. Sentiment analysis of online users' negative emotions based on graph convolutional network and dependency parsing[J]. Data Analysis and Knowledge Discovery，2021，5(9)：97-106. (in Chinese)

[11] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/1301. 3781v3.

[12] 方红，苏铭，冯一铂，等. 结合 gazetteers 和句法依存树的中文命名实体识别[J]. 计算机工程与应用，2022，58(18)：227-232.
FANG H，SU M，FENG Y B，et al. Chinese named entity recognition combined with gazetteers and syntactic dependency tree[J]. Computer Engineering and Applications，2022，58(18)：227-232. (in Chinese)

[13] 张军莲，张一帆，汪鸣泉. 基于图卷积神经网络的中文实体关系联合抽取[J]. 计算机工程，2021，47(12)：103-111.
ZHANG J L，ZHANG Y F，WANG M Q，et al. Joint extraction of Chinese entity relations based on graph convolutional neural network[J]. Computer Engineering，2021，47(12)：103-111. (in Chinese)

[14] LI Y Z，YU B W，XUE M G，et al. Enhancing pre-trained Chinese character representation with word-aligned attention[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/1911. 02821v2.

[15] VASHISHTH S，BHANDARI M，YADAV P，et al. Incorporating syntactic and semantic information in word embeddings using graph convolutional networks[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/1809. 04283v4.

[16] HOU X C，QI P，WANG G T，et al. Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/2103. 11794v1.

[17] SALTON G. The SMART retrieval system：experiments in automatic document processing[M]. Englewood Cliffs，NJ：Prentice-Hall，1971.

[18] 余传明，薛浩东，江一帆. 基于深度交互的文本匹配模型研究[J]. 情报学报，2021，40(10)：1015-1026.
YU C M，XUE H D，JIANG Y F. Research on text matching model based on deep interaction[J]. Journal of the China Society for Scientific and Technical Information，2021，40(10)：1015-1026. (in Chinese)

[19] PANG L，LAN Y Y，GUO J F，et al. Text matching as image recognition[EB/OL]. [2023-10-05]. https：//arxiv. org/abs/1602. 06359v1.

[20] HUANG P S，HE X D，GAO J F，et al. Learning deep structured semantic models for web search using clickthrough data[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York，USA：ACM Press，2013：2333-2338.

[21] REIMERS N，GUREVYCH I. Sentence-BERT：sentence embeddings using Siamese-BERT-networks[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/1908. 10084v1.

[22] CHE W X，FENG Y L，QIN L B，et al. N-LTP：an open-source neural language technology platform for Chinese[EB/OL]. [2023-10-05]. http：//arxiv. org/abs/2009. 11616v4.

[23] VASWANI A，SHAZEER N，PARMAR N，et al. Attention

计算机工程 www.ecice.com

is all you need[EB/OL]. [2023-10-05]. https://arxiv.org/abs/1706.03762.
[24] WANG Z G, HAMZA W, FLORIAN R. Bilateral multi-perspective matching for natural language sentences[EB/OL]. [2023-10-05]. http://arxiv.org/abs/1702.03814v3.
[25] CHEN Q, ZHU X D, LING Z H, et al. Enhanced LSTM for natural language inference[EB/OL]. [2023-10-05]. http://arxiv.org/abs/1609.06038v3.
[26] YANG R Q, ZHANG J H, GAO X, et al. Simple and effective text matching with richer alignment features[EB/OL]. [2023-10-05]. http://arxiv.org/abs/1908.00300v1.
[27] SU J, CAO J, LIU W, et al. Whitening sentence representations for better semantics and faster retrieval[EB/OL]. [2023-10-05]. https://arxiv.org/abs/2103.15316?context=cs.CL.
[28] SU J, AHMED M, LIU B Y. RoFormer: enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568(1):127063.
[29] SUN Y, WANG S H, LI Y K, et al. ERNIE 2.0: a continual pre-training framework for language understanding[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8968-8975.
[30] SUN Y, WANG S H, FENG S K, et al. ERNIE 3.0: large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL]. [2023-10-05]. http://arxiv.org/abs/2107.02137v1.

编辑 吴云芳

计算机工程 www.ecice06.com