



计算机科学

COMPUTER SCIENCE

融合多头注意力机制和孪生网络的语义匹配方法

臧洁, 周万林, 王妍

引用本文

臧洁, 周万林, 王妍. 融合多头注意力机制和孪生网络的语义匹配方法[J]. 计算机科学, 2023, 50(12): 294-301.

ZANG Jie, ZHOU Wanlin, WANG Yan. [Semantic Matching Method Integrating Multi-head Attention Mechanism and Siamese Network](#) [J]. Computer Science, 2023, 50(12): 294-301.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于自适应搜索范围调整的视觉目标跟踪](#)

Visual Object Tracking Based on Adaptive Search Range Adjustment

计算机科学, 2023, 50(11A): 221000172-6. <https://doi.org/10.11896/jsjcx.221000172>

[融合机器阅读理解的中文医学命名实体识别方法](#)

Chinese Medical Named Entity Recognition Method Incorporating Machine Reading Comprehension

计算机科学, 2023, 50(9): 287-294. <https://doi.org/10.11896/jsjcx.220900226>

[注意力特征融合的孪生网络目标跟踪方法](#)

Attentional Feature Fusion Approach for Siamese Network Based Object Tracking

计算机科学, 2023, 50(6A): 220300237-9. <https://doi.org/10.11896/jsjcx.220300237>

[基于孪生注意力网络的建设用地遥感影像变化检测](#)

Remote Sensing Image Change Detection of Construction Land Based on Siamese Attention Network

计算机科学, 2023, 50(6A): 220500040-5. <https://doi.org/10.11896/jsjcx.220500040>

[改进MFCC和并行混合模型的语音情感识别](#)

Speech Emotion Recognition Based on Improved MFCC and Parallel Hybrid Model

计算机科学, 2023, 50(6A): 220800211-7. <https://doi.org/10.11896/jsjcx.220800211>

融合多头注意力机制和孪生网络的语义匹配方法

臧洁 周万林 王妍

辽宁大学信息学院 沈阳 110036

(zangjie@lnu.edu.cn)

摘要 考虑企业资源与客户需求匹配问题,现有的方法存在资源和需求封装不够准确以及匹配效果无法满足用户需求等问题。为解决企业资源与需求描述的多样性和歧义性,提出了动态自定义模板封装。针对封装后的需求与资源大多都是中文短文本这一特点,兼顾句子间语义的差异性和相似性,提出了融合多头注意力机制和孪生网络的交互型文本匹配模型。模型使用字词混向量作为输入增强文本的语义信息,将孪生网络与多头注意力机制相融合,作为独立单元提取上下文的语义特征并使语义特征充分交互。为了验证模型的有效性,在经典数据集 LCQMC 和自我构建的 CSMD 数据集上对模型进行了实验,结果表明所提模型在准确率和性能等方面均有不同程度的提升,为企业资源与需求提供了更精准的匹配方法。

关键词: 自定义模板;语义匹配;孪生网络;多头注意力机制;双向 GRU

中图法分类号 TP391

Semantic Matching Method Integrating Multi-head Attention Mechanism and Siamese Network

ZANG Jie, ZHOU Wanlin and WANG Yan

College of Information, Liaoning University, Shenyang 110036, China

Abstract Considering the matching problem of enterprise resources and customer requirements, the existing methods have the problems that the resource and requirement encapsulation is not accurate enough and the matching effect can't satisfy users' requirement. In order to solve the problem of diversity and ambiguity of enterprise resource and requirement description, this paper proposes the dynamic user-defined template encapsulation. According to the feature that most of the encapsulated requirements and resources are Chinese short texts, an interactive text matching model which integrates multi-head attention mechanism and siamese network is proposed. The semantic differences and similarities between sentences are considered in this model. It uses word mixing vectors as input to enhance the semantic information of the text, combines the Siamese network with the multi-head attention mechanism, and extractes the semantic features of the context as an independent unit to fully interact with the semantic features. In order to verify the effectiveness of the model, the classical data set LCQMC and the self-constructed CSMD data set are used to conduct experiments on the model. The results show that the accuracy and performance of the model are improved in different degrees, which provides a more accurate matching method for enterprise resources and requirements.

Keywords Custom template, Semantic matching, Siamese network, Multi-head attention mechanism, Bidirectional GRU

在过去十年中,信息技术应用于各领域,新的数字技术不仅为企业的信息编目、存储和分析提供了强大的工具,而且已成为信息交流的新的通用语言^[1]。不同领域的企业融合也变得越来越流行,这就对资源与需求匹配等方面提出了新的要求。多数情况下,这些企业间的上下游合作通过线下或传统的方式进行,不能很好地解决需求与资源的最优化匹配以及对新企业的安全判断等问题。因此,准确识别和抽象用户需求以及提供可用、有效且匹配的资源池非常重要。

为了提供有效、匹配的可用资源池,首先必须对用户需求进行识别和抽象。现有的识别和抽象方法主要有模板封装法^[2-7]和组件抽象法^[8-11]。模板封装法是针对服务或流程,

创建模板进行封装;组件抽象法是将需求划分为独立的组件,使用组件抽象来表示知识,以达到对问题进行分解的目的。针对多领域需求具有多样性和部分需求高度可重用的特点,以上两种方法分别存在无法对实体资源进行描述、模板对于新手操作难度过大或无法对解决的问题进行保存等缺点,需要提出新的识别和抽象用户需求的方法。

多领域融合的企业对资源的描述具有歧义性,需求也具有多样性和高度可重用的特点,因此提出采用动态的自定义模板封装方法,由用户自己定义资源。资源封装后大多都是中文短文本数据,针对中文短文本匹配方面的研究,主要有文本语义表征模型^[12-17]和文本语义交互模型^[18-22]。文本语义

到稿日期:2022-10-11 返修日期:2023-03-25

基金项目:国家重点研发计划(2019YFB1405804)

This work was supported by the National Key R&D Program of China(2019YFB1405804).

通信作者:王妍(wang_yan@lnu.edu.cn)

表征模型中常用的方法是孪生网络,其采用权重共享的策略,将需要对比的文本分别输入到共享参数的左右神经网络中,得到两组新的向量表示;对新向量进行相似度计算,得到两条文本的相似度,其没有使用文本间的交互信息^[16]。文本语义交互模型首先将需要对比的文本输入到权重共享的左右网络中,计算得到两条文本的特征向量;之后并不直接进行相似度计算,而是对两个句子之间的词向量进行匹配,将匹配的结果输入到一个神经网络再编码;最后进行相似度计算得到文本相似度。但该类方法存在提取交互语义信息维度不够全面、对原始数据集中的噪声问题处理不佳,以及没有很好地利用文本差异性问题。综上,文本语义表征模型中存在句子间的交互信息被忽略的情况,而在文本语义交互模型中存在语义信息提取不够充分和没有很好地利用文本差异性等缺点。

现有的识别、抽象方法和提供的有效匹配池都存在一些不足,不适用多领域融合下企业的资源与需求匹配。针对多领域融合存在的问题,本文首先提出了动态的自定义模板方法来解决资源的抽象和封装;为了提供高效准确的匹配,考虑两个句子间语义的差异性和相似性,提出融合多头注意力机制和孪生网络的交互型文本匹配模型(Integrating Multi-head Attention Mechanism and Siamese Network, IMAMSN)。IMAMSN模型在输入层以字词混合向量作为输入,然后在差异单元和交互单元分别进行处理。在差异单元,对输入的两个字词混合向量采用双向门控循环单元网络(Gated Recurrent Unit, GRU)进行编码,构建孪生网络,然后使用Manhattan距离对编码信息进行聚合。在交互单元,多头注意力机制的自编码层对输入的两个字词混合向量分别进行编码,构建孪生网络,多头注意力机制从不同的角度提取文本的语义信息,孪生网络把序列长度不同的文本调整成长度相同的文本,存在于同一个语义空间;交互层对自编码层的语义特征进行交互,再将交互的信息与差异单元的信息进行融合。最后,通过两个实验对模型算法进行了验证,结果验证了模型的正确性和有效性。

1 相关工作

1.1 资源封装

现有的资源封装技术主要分为模板封装法和组件抽象法。Chen等^[2]根据服务提供方式对制造资源进行分类,建立基于Web服务的云制造软资源封装模型。但由于每个软件资源都要根据预先定义好的属性和接口描述文档(Web Service Description Language, WSDL文档)进行接口类的实现,当模板不能满足资源的属性和功能接口的要求时,需要用户自定义属性描述和接口实现,这对于新手用户来说使用难度增大且并不友好。为了提高地表水环境决策支持系统(Environmental Decision Support System, EDSS)开发的效率并降低其复杂性,Jiang等^[3]提出了一种将地表水水质模型及其应用封装为云计算服务的通用途径。Kong等^[4]提出了基于服务模板的制造资源封装方法,利用已有服务程序进行模板抽取,充分利用已有的制造资源服务程序,加快制造资源封装进程。为了解决用户希望通过在三维虚拟世界中定制实时

Metaverse服务来获得沉浸式和交互式体验,却在服务的管理和编排方面没有统一解决方案的问题,Liu等^[5]提出了名为元宇宙服务(Metaverse as a Service, MaaS)的解决方法,把Metaverse中的各种物理虚拟组件和技术封装为服务交付。Luckow等^[6]针对许多科学和工业物联网应用中边缘到云的数据处理需求,提出了Pilot Edge作为跨边缘到云连续体的资源管理通用抽象,通过Pilot Edge从应用程序中提取资源。Pawlewski等^[7]使用关系数据库模板在制造系统中进行建模,使用每个零件计划(Plan for Every Part, PEEP)作为模型的主要集成器。

组件抽象法^[8]将需求划分为独立的组件,并使用组件抽象来表示知识,使用启发式规则分解给定的问题,以便在组件级别进行规划。Antequera等^[9]以组件抽象模型为特色,以“自定义模板”的形式实现智能资源“抽象”,再加上“可重用”的硬件和软件配置,以简化异构资源管理工作。Chen等^[10]提出了一种模型知识封装方法,解决了钢铁调度生产模型的快速重用和动态构建问题。该方法由知识封装框架和知识映射方法组成,知识封装框架旨在帮助模型提供者在开放网络环境中封装调度模型,构造映射方法以加强模型知识之间的关系。Tian等^[11]提出了一种基于终端弹性可重构应用服务架构的轻量化增量资源封装方法,采用虚拟资源盒来规划展示各个任务单元的已有资源和需要装载的增量资源。

综上,现有方法存在无法对实体资源进行封装、操作难度过大、不利于新手的使用以及静态模板很少与用户进行交互等问题。因此,本文提出动态自定义模板以解决需求与资源抽象到资源组件中的问题,增加了更多的参数对特征进行描述,使问卷系统的交互更加动态,类似于现实中“用户-专家”之间进行交互。

1.2 文本语义匹配模型

针对语义匹配问题,目前主要提出了两种模型,一是文本语义表征模型,二是文本语义交互模型。文本语义表征模型关注于文本表示,将待匹配的两个文本极力地表示成相同长度的语义表示向量,然后进行匹配度的计算。Huang等^[12]提出了深度语义结构模型(Deep Structured Semantic Model, DSSM),他们最早在深度文本匹配模型中应用孪生网络,采用全连接网络进行编码,最后计算匹配文本的余弦相似度。Wang等^[13]提出基于孪生网络的多粒度匹配模型,在句子的深层语义相似度和浅层语义相似度上充分挖掘句子间的相似信息。Mueller等^[14]使用双向长短期记忆网络(Long Short-Term Memory, LSTM)分别提取前向和后向的上下文信息并构造孪生网络,使用余弦相似度来计算文本相似度。Deng等^[15]提出了一种融合多神经网络的文本分类模型(A Text Classification Model Fused with Multiple Neural Network, FMNN),采用不同的方式提取全局语义特征和局部语义特征,充分利用了各模型的优势,提取多方面的语义特征来保留更多的语义信息。Deng等^[16]针对多头注意力机制很难进一步获取更深层次的内联关系的问题,提出了一种短文本情感分类模型,利用改进的多头注意力机制和多跳记忆力网络对短文本上下文内部关联进行有效建模,充分挖掘高层情感

语义特征。Zhang 等^[17]提出了一种新的多粒度融合的句子语义匹配模型,从字符和单词角度获得文本的语义特征,并通过这两个特征的叠加效应捕获文本中更关键的语义信息,并设计了一种新的损失函数,利用均方误差(Mean Square Error, MSE)作为交叉熵的平衡因子增强区分模糊分类边界的能力。文献[15-17]以不同的方式获得丰富的语义特征,并根据自己的特色挖掘高层情感语义特征或引入新的损失函数,但它们都属于文本语义表征模型,没有充分关注特征之间的交互融合。

文本之间的交互对于文本匹配是十分重要的,因此研究者们提出了第二种文本语义交互模型。交互型深度匹配模型侧重文本之间的交互,获取到的深层语义特征更加丰富,角度更加多样。Chen 等^[18]提出了(Enhanced Sequential Inference Model, ESIM)模型,该模型主要分为编码、交互和比较 3 部分,首先采用双层双向的 LSTM 对输入的两条文本进行编码,获得两条新的编码序列;对于文本语义表征模型没有特征间交互的问题,引入了注意力机制学习文本间的交互信息,同时将树 LSTM 编码序列和交互特征序列拼接起来做特征增强,最终输入到定制化的相似度计算模块。ESIM 虽然采用树 LSTM 来帮助收集短句与子句之间的文本信息,但是对于短文本来说收集到的信息还是不够充分。Wu 等^[19]对同一短文本采用不同粒度的语义表达,通过两个文本不同粒度的语义表示交互获取两个短文本的交互特征。Chen 等^[20]提出了自监督模型的辅助任务,以句子对中两个句子相互生成的方式,获取文本之间基于序列转换的深层交互信息,以特征增强的方式辅助下游模型。文献[19]和文献[20]以不同方式获取更多短文本交互特征,来提高模型的准确性。Ma 等^[21]研究了短文本匹配中模型在不同数据集之间的泛化能力,以及如何在新的领域有效利用现有不同领域中的大量带标注数据,达到减少新领域数据标注量的目的。Xu 等^[22]提出了一种融合实体上下文特征的深度文本语义匹配模型,兼顾了文本匹配过程中实体上下文和语义关联信息,该模型通过学习深度多视图语义交互信息和实体上下文特征匹配矩阵来计算文本的综合匹配得分。

在上述的语义匹配算法中,既有文本语义表征算法,又有文本语义交互算法,虽然都进行了较好的匹配,但仍仍然存在一些问题。在文本语义表征算法中,没有关注两个句子之间交互信息的利用;在文本语义交互算法中,没有充分提取文本的语义特征且忽略了文本差异性。针对以上问题,本文提出了融合多头注意力机制和孪生网络的交互型文本匹配模型,使得句子对的匹配过程兼顾差异性与相似性,充分提取了文本的语义特征。

2 动态的自定义模板资源封装

由于多领域融合后的资源具有多样性和歧义性,同时,需求本身也具有多样性和部分需求具有高可重用性。针对这些特征,考虑使用动态自定义模板对资源进行封装,由用户自行定义资源的类型和描述。动态自定义模板的示意图如图 1 所示。

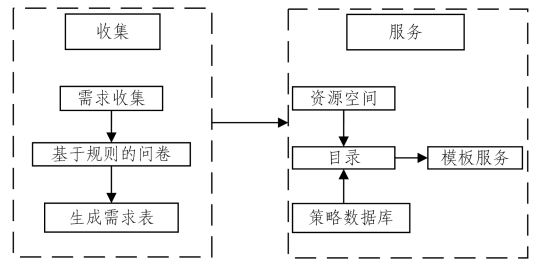


图 1 动态自定义模板

Fig. 1 Dynamic custom template

动态自定义模板主要分为收集和服务两阶段,收集阶段为服务阶段做信息收集准备。

1) 收集:这一阶段通过用户需求收集系统(User Demand Collection System, UDCS)收集和抽象用户的需求。UDCS 是一个基于 Web 的解决方案,通过模拟场景来收集融合公司的需求。通常,需求收集是通过用户和公司业务员之间的一对一访谈和会议方式完成的,在动态自定义模板中,则是通过向用户提供一份基于规则的调查问卷,以收集用户需求,从而生成需求表。

2) 服务:其中资源空间用来提供模板匹配服务;目录保存之前应用过的解决方案,以方便用户下一次使用;策略数据库提供政策规定与先验知识,如指导用户某一个需求必须在另一个需求之前填写。

当用户完成与 UDCS 的交互时,收集的数据自动转换为用户需求标识符(User demand identifier, UDI)结构。通过自定义模板收集数据,自定义模板的每个模块与 UDI 中的每个域是对应关系,模块中的问题对应域中的资源描述特征。UDI 结构表示用户需求的元信息,其中包含一个向量 U , 表示为 $U = \{U_1, U_2, \dots, U_K\}$, $U_K = \{F_1 - P_1, F_2 - P_2, \dots, F_i - P_i\}$ 。 U 是一个用户需求标识符,一个需求标识符被划分为 K 个领域, U_K 由满足用户需求的资源特征和相应的前提条件组成,每个领域由 i 个 $F - P$ 特征进行描述,其中 F 代表用户需求的特征, P 表示 F 特征相应的前提条件。

资源空间被分为多个域,比如文化、电竞和科技等领域。本文定义 S 是所有领域资源的集合,其中 $S = \{D_1, D_2, \dots, D_N\}$, N 等于域的数量。每个这样的域由不同数量的特征领域组成, $D_N = \{T_1^N, T_2^N, \dots, T_N^N\}$, T_1^N 到 T_N^N 表示 D_N 域中不同的特征领域,每个特征领域由 R_N 个资源组成,表示为 $T_N^N = \{R_1^N, R_2^N, \dots, R_N^N\}$ 。例如在科技领域中,健康特征领域下部分企业数据主营业务部分的展示,如表 1 所列。

表 1 部分企业数据主营业务的展示

Table 1 Display of main business of some enterprise data

统一社会信用代码	企业名称	企业位置	主营业务
91340100610307675N	融捷健康科技股份有限公司	安徽省合肥市长江西路 669 号 AJ-9 地块	远红外线桑拿等衍生设备、机电产品的生产、委托加工、销售及自营进出口业务(凭许可证经营);医疗器械二类:6826 物理治疗及康复设备的生产、销售;木材、木制品、建筑装潢材料、电子产品及其他产品的销售

本文通过 UDCS 收集和抽象用户的需求,然后将收集到

的数据与资源空间中的资源进行比较。由需求表示符划分的 K 个领域对应资源空间中领域的类别数量,用户需求的资源特征对应资源空间中的特征领域。因此,满足条件 k 的值小于 D_N 中 N 的值,满足条件 i 的值小于资源空间中 T_N^A 中 A 的值。

3 资源与需求的匹配模型

由于封装后的资源与需求大多都是中文短文本,针对该特点,提出了融合多头注意力机制和孪生网络的交互型文本匹配模型。模型中采用字词混合向量作为输入来增加语义信息,使用 Bi-GRU 代替 Bi-LSTM 来减少参数并加快训练速度;将多头注意力机制作为单独的模块使用,在自编码层使用多头注意力自注意力机制从不同的角度获取文本的语义特征,并构建孪生网络使长度不同的句子变成长度相同的句子,存在于同一个语义空间,同时,权重共享可以减少一半的训练量。在交互层,将句子的字词混合向量与另一个句子的字词混合向量进行交互,使用多头注意力机制来获取交互的语义特征。

本章首先介绍孪生神经网络模型,在此基础上详述提出的融合多头注意力机制和孪生网络的交互型文本匹配模型。

3.1 孪生神经网络模型

孪生神经网络模型示意图如图 2 所示。

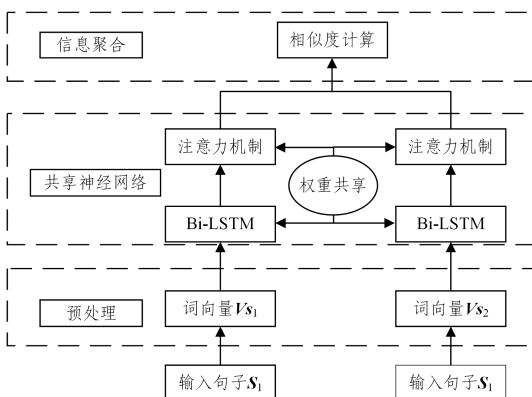


图 2 孪生网络模型

Fig. 2 Siamese network model

孪生网络模型大致分为预处理、共享神经网络和信息聚合 3 部分。孪生网络的处理流程是:首先将获得的句子进行预处理,得到句子的词向量表示;然后将得到的句子表示由 Bi-LSTM 和注意力机制构建的孪生网络进行编码;最后对经过孪生网络的句子表示进行信息聚合。

1) 预处理

首先获得包含上下文信息的句子,用词向量表示句子的每个词语。句子表示为 $V_{S_1} \in R^{m \times d}$, $V_{S_2} \in R^{n \times d}$ 。其中 m 和 n 代表句子 S_1 和 S_2 的长度, d 为字词向量的维度。

2) 共享神经网络

在得到句子的表示后,首先会经过一个 Bi-LSTM,通过 Bi-LSTM 可以得到句子向量的编码信息,这里使用标准的 Bi-LSTM,其公式如式(1)所示:

$$\begin{cases} \vec{h}_t = f(W_1 x_t + W_3 \vec{h}_{t-1} + \vec{b}_t) \\ \overleftarrow{h}_t = f(W_2 x_t, W_4 \overleftarrow{h}_{t+1} + \vec{b}_t) \\ y_t = \text{concat}(\vec{h}_t, \overleftarrow{h}_t) \end{cases} \quad (1)$$

其中, $\vec{h}_t, \overleftarrow{h}_t, x_t, y_t$ 分别表示在 t 时刻的正向传播隐藏层状态、反向传播隐藏层状态、输入神经元的输入值和隐藏层状态的输出值。 \vec{h}_{t-1} 和 \overleftarrow{h}_{t+1} 表示在 $t-1$ 时刻的正向传播状态和 $t+1$ 时刻的反向传播状态, W_1, W_2, W_3 和 W_4 分别表示不同组件对应的权重矩阵。

在 Bi-LSTM 层之后加入注意力机制,共享权重的神经网络包含上面的 Bi-LSTM 层以及注意力机制层。

3) 信息聚合

处理后的两个句子需要进行信息融合,常见的融合方式包括全连接神经网络、计算两个向量的余弦相似度和 Manhattan 距离等。

3.2 融合多头注意力机制和孪生网络的交互型文本匹配模型

孪生网络模型没有充分地利用文本间的交互信息,也未能更好地获得文本的表征能力,因此,本文对现有的孪生网络模型进行了改进,提出了融合多头注意力机制和孪生网络的交互型文本匹配模型,如图 3 所示。

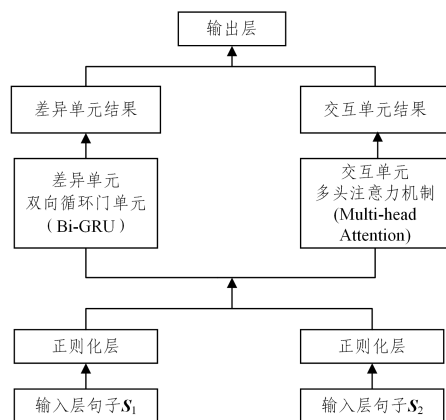


图 3 融合多头注意力机制和孪生网络的交互型文本匹配模型

Fig. 3 Interactive text matching model integrating multi-head attention mechanism and siamese network

该模型首先在输入层对两个句子进行预处理,得到两个句子的字词混合向量,将得到的字词混合向量经过正则化层进行正则化;然后把处理后的向量分别输入到差异单元和交互单元。在差异单元,首先对输入的两个文本采用双向 GRU 方式进行编码、构建孪生网络,然后使用 Manhattan 距离对编码信息进行聚合。同时,在交互单元,首先利用多头注意力机制的自编码层对输入的两个句子分别进行编码并构建孪生网络,然后在多头注意力机制的交互层对自编码层的语义特征进行交互,最后将交互的信息与差异单元的信息作为结果输出到输出层的全连接神经网络中,经过 sigmoid 获得分类结果。

3.2.1 字词混合向量

现有的模型大多采用字向量或词向量输入,对于中文数据集来说,字和词都携带了大量的语义信息,单独使用任何

一种都无法很好地表达句子信息。因此,考虑使用字词混合向量代替字向量和词向量,或将字向量和词向量分别输入以提高句子的语义表征能力。

首先用 jieba 分词^[23]工具对训练集进行词和字符级别分割,对于输入的句子 S_1 和 S_2 ,经过词和字符级别分割后,可以得到词序列和字符序列两种表示。通过加载预训练 Word2vec 模型的权重,可将每个单词或字符映射到一个向量中,提取对应的字词向量,把词向量的长度扩充为字向量的长度。将融合的字词混合向量 V_{S_1} 和 V_{S_2} 输入到正则化层。对句子 S_i 的形式化描述如式(2)所示:

$$\begin{aligned} S_i &= \{W_1, W_2, \dots, W_m\} \\ S_i &= \{C_1, C_2, \dots, C_m\} \end{aligned} \quad (2)$$

其中, W_m 和 C_m 分别是长度为 m 的句子 S_i 的词向量表示和字向量表示。

3.2.2 采用 Bi-GRU 结构替代 Bi-LSTM

本文采用 Bi-GRU 代替孪生网络中的 Bi-LSTM 进行文本信息编码。GRU^[24]是由 Cho 等在 2014 年提出的 LSTM 变种循环门单元(Gated Recurrent Unit, GRU),该模型用一个更新门代替 LSTM 中的遗忘门和输入门,并且定义了新的隐藏单元,模型结构比 LSTM 更简单,计算式如式(3)和式(4)所示:

$$\begin{cases} z_t = \sigma(w_z \cdot [h_{t-1}, x_t] + b_z) \\ r_t = \sigma(w_r \cdot [h_{t-1}, x_t] + b_r) \\ \tilde{h}_t = \tanh(w_h \cdot [r_t \cdot h_{t-1}, x_t] + b_h) \end{cases} \quad (3)$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \quad (4)$$

其中, z_t 为更新门, r_t 为重置门。更新门表示新的隐藏状态继承了多少当前的隐藏状态;重置门表示过去的隐藏状态中有多少信息是可被忽略的,并使用当前输入进行重置。GRU 的参数比 LSTM 少,训练的速度更快,因此所需要的数据更少。

3.2.3 交互单元

孪生网络模型在神经网络后面直接引入注意力机制,提取句子的重要词信息,没有把句子对的语义特征进行充分的提取并进行交互。因此,本文把多头注意力机制单独作为一个单元进行计算,提取文本对之间的交互特征,将基础的注意力网络分割成不同的子网络,从多个不同的角度学习更多的语义特征,达到语义特征充分交互的目的。交互单元示意图如图 4 所示。

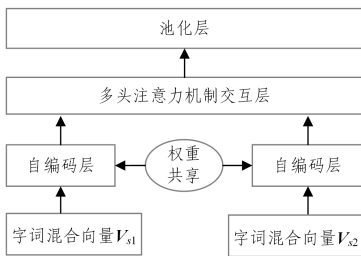


图 4 交互单元示意图

Fig. 4 Schematic diagram of interaction unit

注意力交互单元分为自编码层、交互层和池化层。

1)自编码层:在自编码层对输入的字词混合向量 V_{S_1} 和 V_{S_2} 分别进行编码。自编码层由两部分组成,即多头注意力层和前向网络。多头注意力层从不同的维度获取文本中的语义信息,前向神经网络通过加入非线性变化来提高特征的提取能力,在多头注意力层和前向网络后面都加入层正则化和残差连接以提升模型的性能,加快模型的训练速度。

V_{S_i} 经过 3 次线性变化分别得到 *Query* 矩阵、*Key* 矩阵和 *Value* 矩阵。之后,采用缩放点积注意力计算自注意力关联结果,注意力机制相当于在查询过程中抽取语义特征并进行编码,使每个词拥有 3 个向量,一次查询操作类似于对一个单词的 *Query* 向量与全部 *key* 向量进行点乘计算相似度,得到的结果即为单词与其他词的权重系数,同时用权重系数与全部的 *value* 向量做加权得到语义编码。最后,连接所有的多头自注意力机制的输出,并输入到带有 Relu 激活函数的前向网络中。整个过程如式(5)~式(8)所示:

$$\begin{cases} Q_i^v = V_{S_1}^i S_1^v \\ K_i^k = V_{S_2}^i S_2^k \\ V_i^s = V_{S_1}^i S_1^s \end{cases} \quad (5)$$

$$Att_i^v = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_q}}\right) V_i^s \quad (6)$$

$$\begin{cases} Ma = [Att_i^v, \dots, Att_i^v] \\ Ma = \text{LayerNorm}(Ma) \end{cases} \quad (7)$$

$$\begin{cases} Ha = \text{relu}(Ma W_1^v) W_2^s \\ Ha = \text{LayerNorm}(Ha) \end{cases} \quad (8)$$

2)交互层:在编码层中,对文本分别进行编码,获得了多样的上下文语义信息,使文本特征能够更好地进行文本表示,但由于孪生网络极力将长度不同的语义特征向量在同一语义空间进行表示,导致文本间的交互信息被忽略。与普通的 Web 文本相比,用户资源和需求文本都是长度较短的短句子,且句子之间并没有上下文语义关系。因此引入交互层,分别提取文本对之间的词句交互特征,每个词语都能与另一个文本中的词语进行交互,得到文本对之间的句法、语义依赖关系,用来弥补孪生网络没有进行语义交互的缺陷。利用注意力机制对两条文本的交互信息进行获取,其公式如式(9)和式(10)所示:

$$\begin{cases} \tilde{U} = \{\tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_{L-1}\} \\ \tilde{V} = \{\tilde{v}_0, \tilde{v}_1, \dots, \tilde{v}_{L-1}\} \end{cases} \quad (9)$$

$$\begin{cases} \tilde{u}_i = \text{MultiHead}(u_i, V, V) \\ \tilde{v}_i = \text{MultiHead}(v_i, U, U) \\ \text{MultiHead}(\text{Query}, \text{Key}, \text{Value}) = \\ \quad \text{Concat}(\text{head}_1, \dots, \text{head}_8) W^0 \\ \text{head}_i = \text{Attention}(\text{Query}_i, \text{Key}_i, \text{Value}_i) \\ \quad i = 1, \dots, 8 \\ \text{Attention}(\text{Query}, \text{Key}, \text{Value}) = \\ \quad \text{softmax}\left(\frac{\text{Query} \cdot \text{Key}^T}{\sqrt{d_k}}\right) \text{Value} \end{cases} \quad (10)$$

3)池化层:在池化层采用全局平均池化来精炼位置向量的特征。

输出时,将得到的交互单元结果和差异单元结果进行融合,通过一个全连接神经网络,经过 sigmoid 获得分类结果。形式化描述为式(11)和式(12):

$$Y = \text{Sigmoid}(FC(\mathbf{X}_{\text{差异}}, \mathbf{X}_{\text{交互}})) \quad (11)$$

$$FC(\mathbf{X}) = \text{relu}(\mathbf{w} \cdot \mathbf{x} + b) \quad (12)$$

模型采用均方误差作为损失函数,如式(13)所示。同时,使用 Adam 优化算法提高收敛速度。

$$\text{LossMSE} = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 \quad (13)$$

4 实验与结果分析

为了充分证明融合多头注意力机制和孪生网络的交互型文本匹配模型的有效性,本文进行了两部分实验,第一部分采用经典数据集(LCQMC)^[25]进行训练并测试,第二部分采用自我构建的文化语义匹配数据集(CSMD)进行测试,与自定义推荐模板算法进行对比,验证模型在需求与资源匹配上的效果。

4.1 实验数据与实验设置

4.1.1 数据集

第一部分实验采用的数据集是中文问题语义匹配数据集(LCQMC),由于资源与需求文本是中文短文本数据并且各个问题之间相关性并不高,与 LCQMC 数据集高度相似,所以第一部分采用 LCQMC 数据集训练并验证模型。LCQMC 数据集更侧重于意图匹配,首先使用搜索引擎从各个领域收集与高频词相关的大规模问题对,然后通过 Wasserstein 距离过滤不相关的问题对,最后通过手工进行验证,因此,这是一个高质量的数据集。数据集一共由 260068 个问题对构成,其中训练集问题对 238766 个,验证集问题对 8802 个,测试集问题对 12500 个。

第二部分实验采用的数据集是文化语义匹配数据集(CSMD),这是根据项目需求,由动态自定义模板收集到的用户文化需求与资源空间的文化资源。首先采用 K-means 算法对收集到的用户需求 and 资源空间中的资源根据特征领域进行聚类,然后使用 Wasserstein 距离排除质量不高的匹配对,保留 Wasserstein 距离在 0.15~0.45 之间的数据,然后对剩下的这部分数据进行手工标注。构建出的文化语义匹配数据集,包含 4000 多条测试数据,匹配和不匹配的比例大约为 7:3。对于不匹配的数据,尽量采用 Wasserstein 距离越小的数据,较小的距离与较高重叠的词汇在同一句子对中,仅有几个词语不同,对于算法来说很难识别。

4.1.2 实验设置

本文既用到词向量又用到字向量,因此,分别使用 Word2Vec 训练文本中字和词的向量表示,每个向量的维度都是 128 维。在实验中,为了缓解过拟合、提高准确率,加入了摒弃舍弃概率 Dropout。通过输入层后,句子向量被输入到双层双向的 GRU 中,两个 GRU 的隐藏层维度均设置为 128。注意力机制共有 8 个,激活函数采用 Relu。为了训练模型,实验采用了 Adam 优化器,学习率设置为 10^{-3} ;为了取得更好的实验效果,本文使用了早停的训练方法。

为了选取合适的 Dropout 值,本文进行了不同 Dropout 值的对比实验。以 F1 值为评估指标,在 LCQMC 数据集上进行验证,实验结果如图 5 所示。由图可知,Dropout 值在 0.1 和 0.8 时 F1 值较低,在 Dropout 值为 0.5 时 F1 值达到最高,因此本文采用的 Dropout 值为 0.5。

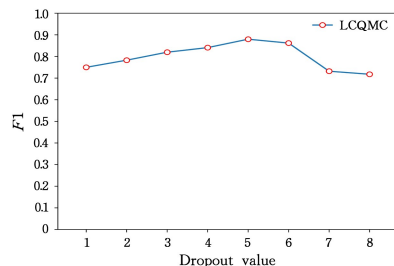


图 5 不同 Dropout 值对比实验

Fig. 5 Comparison experiment of different Dropout values

4.2 LCQMC 实验结果与分析

为了验证本文模型(IMAMSN)的有效性,将结果与 4 个不同的模型进行了对比,并采用准确率、召回率、精确率以及 F1 值对模型进行评价。

卷积神经网络(CNN)^[26]:输入的每个文本表示为一个词向量矩阵,该矩阵通过一个卷积神经网络进行信息聚合,然后再进行相似度的计算。

Siamese LSTM^[27]:将两个句子分别输入到双向双层 LSTM 网络中,双向 LSTM 可以获取前后的上下文信息,双层 LSTM 可以学习到更多的语义特征。两个句子经过权重共享的神经网络模块,得到同一维度的两条特征向量,因此相似度计算采用余弦相似度。

双边多视角匹配(BiMPPM)^[28]:BiMPPM 使用基于字符的 LSTM 在单词表示层构建不同的多视角匹配操作,在不同视野下应用相似度函数提取不同的交互特征,最后输入两层前馈网络进行预测。

孪生网络和 Am-softmax 损失函数(SNAS)^[29]:SNAS 在其输入层分别输入字向量和词向量进行训练,采用 BiLSTM 获取文本语义信息,之后引入注意力机制提取对句子含义重要的词信息。实验结果如表 2 所列。

表 2 LCQMC 数据集实验结果对比

Table 2 Comparison of LCQMC experimental results

Name	Accuracy	Precision	Recall	F1
CNN	0.720	0.673	0.802	0.732
SiameseLSTM	0.774	0.742	0.886	0.808
BiMPPM	0.830	0.769	0.928	0.841
SNAS	0.859	0.798	0.937	0.862
IMAMSN	0.883	0.825	0.922	0.870

从表 2 中可以看出,本文提出的方法的准确率为 0.883,精确率为 0.825,召回率为 0.922,F1 值达到 0.870,在准确率、精确率和 F1 值这 3 个指标上均优于其他模型。

与对比模型相比,本文提出的 IMAMSN 模型在 Embedding 层面使用字词混合向量对句子特征进行了更好的表示。与 CNN 模型相比,本文模型在准确率和 F1 上分别提高了 0.163 和 0.138,证明采用双向 GRU 能够捕获上下文语义

信息,更适合文本序列,能够处理长期依赖的问题。与 Siamese LSTM 模型相比,本文模型在准确率和 F1 方面分别提高了 0.109 和 0.072,这是因为本文模型引入了 Multihead-attention 机制,提取了更多的模型交互特征。与 BiMPM 模型相比,本文模型在准确率和 F1 上分别提高了 0.053 和 0.039,原因在于本文模型使用的双层双向 GRU 训练参数更少,模型结构比 LSTM 更简单。与 SNAS 相比,本文模型在准确率和 F1 方面分别提高了 0.024 和 0.008,原因在于本文将注意力机制单独作为一个模块进行,能够更好地提取句子间的交互特征,且注意力机制的编码层采用孪生网络,使特征有极强的文本表征能力。

图 6 给出了经典模型和本文提出的 IMAMSN 模型的召回率值。从表 2 和图 6 可以发现,IMAMSN 的召回率值比 BiMPM 和 SNAS 的召回率值分别低 0.006 和 0.015,原因是本文模型采用孪生网络和多头注意力机制,充分关注了文本的差异性;与纯文本语义交互算法相比,增大了将正类预测为负类的概率;FN 的增加,导致了召回率的降低。

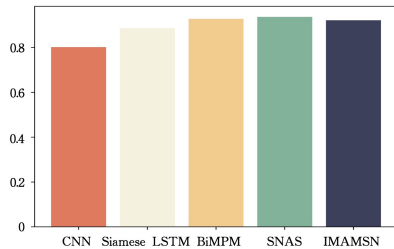


图 6 不同模型的 Recall 值

Fig. 6 Recall values of different models

4.3 CSMD 实验结果与分析

为了进一步验证本文模型在资源与需求匹配上的效果,证明动态自定义模板匹配效果优于原有的自定义模板,本文采用某项目自采数据集(CSMD),将本文方法与自定义模板推荐算法(Custom Template Recommendation Algorithm, CTRA)^[9]进行了对比实验,实验结果如图 7 所示。

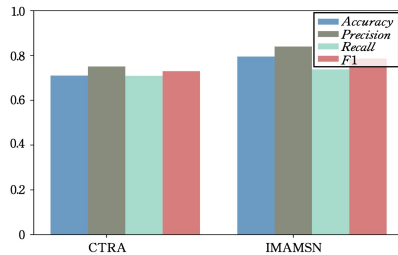


图 7 CSMD 上的对比实验

Fig. 7 Comparative experiment on CSMD

从图 7 中可以看出,本文提出的模型在准确率、精确率、召回率和 F1 值上相比 CTRA 算法均有所提高。其中 CTRA 的准确率和 F1 值分别为 0.710 和 0.730,本文提出的算法准确率和 F1 值分别为 0.795 和 0.786,分别提高了 0.085 和 0.056。CTRA 是将用户表示中的一个特性直接与资源空间中的一个资源或多个资源进行匹配,本文方法较 CTRA 更多地考虑不同领域融合导致资源描述之间的歧义性,兼顾了句子语义之间的差异性与相似性,因此匹配的准确率和

F1 值优于 CTRA。

本文通过两个实验,证明了融合多头注意力机制和孪生网络的交互式文本匹配模型 IMAMSN 进一步提高了文本匹配的准确度,也验证了模型的有效性。

结束语 本文提出一种针对多领域融合中的企业资源与需求匹配方法。多领域企业的融合,导致融合后的资源具有多样性和歧义性,并且需求也具有多样性和部分需求具有高可重用性的特性。针对以上特点,提出了动态自定义模板的封装方法来解决识别、抽象用户需求和资源描述的问题,提出融合多头注意力机制和孪生网络的交互式文本匹配模型,提高了中文短文本匹配的性能。在 LCQMC 和 CSMD 数据集上的实验证明了所提模型进一步提高了文本匹配准确度,同时验证了模型的有效性。

一般来说,生活中产生的大多数数据集都是稀疏且低质量的数据集,模型如何从稀疏的数据集学习到更多语义特征,以及如何对从低质量数据集学习到的有限语义特征做增强处理是未来工作的重心。

参考文献

- [1] CAMPANA S, FRANCOVICH R. Landscape Archaeology in Tuscany: Cultural resource management, remotely sensed techniques, GIS based data integration and interpretation[J]. Bar International Series, 2003, 1151: 15-28.
- [2] CHEN K, WANG D, WANG Y H, et al. Research on Soft Resource Packaging of Cloud Manufacturing [J]. Informatization of China's Manufacturing Industry, 2012, 41(5): 58-63.
- [3] JIANG J, PANG T, ZHANG F, et al. Pathway to encapsulate the surface water quality model and its applications as cloud computing services and integration with EDSS for managing urban water environments[J]. Environmental Modelling & Software, 2022, 148: 105280.
- [4] KONG L J, XU W S, ZHA J Z. Manufacturing resource packaging method based on service template [J]. Computer Applications, 2012, 32(12): 3534-3539.
- [5] LIU Y J, DU H, NIYATO D, et al. Slicing4Meta: An Intelligent Integration Framework with Multi-dimensional Network Resources for Metaverse-as-a-Service in Web 3.0[J]. arXiv: 2208.06081, 2022.
- [6] LUCKOW A, RATTAN K, JHA S. Pilot-Edge: Distributed Resource Management Along the Edge-to-Cloud Continuum[C]// 2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). IEEE, 2021: 874-878.
- [7] PAWLEWSKI P, ANHOLCER M. Relational database template in the simulation modeling of manufacturing systems[J]. IFAC-Papers OnLine, 2019, 52(13): 1744-1748.
- [8] BOTEVA A, MULLER M, SCHAEFFER J. Using Component abstraction for automatic generation of macro-actions [C]// ICAPS. 2004: 181-190.
- [9] ANTEQUERA R B, CALYAM P, CHANDRASHEKARA A A, et al. Recommending heterogeneous resources for science gateway applications based on custom templates composition [J]. Future Generation Computer Systems, 2019, 100: 281-297.

- [10] CHEN X,JIANG G,LI G,et al. A new knowledgeable encapsulation method of steel production scheduling model[J]. Proceedings of the Institution of Mechanical Engineers,Part B:Journal of Engineering Manufacture,2020,234(14):1673-1685.
- [11] TIAN Y,QIU L Y,WU Y F,et al. Microservice based Space based Information Support Tactical Terminal Application Service Architecture [J]. Computer and Network,2021,47(16):64-68.
- [12] HUANG P S,HE X,GAO J,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,2013:2333-2338.
- [13] WANG X,YANG H. MGMSN:Multi-Granularity Matching Model Based on Siamese Neural Network[J]. Frontiers in Bioengineering and Biotechnology,2022,10:839586.
- [14] MUELLER J,THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2016.
- [15] DENG W B,ZHU K,LI Y B,et al. FMNN:Text Classification Model Based on Multi neural Network [J]. Computer Science,2022,49(3):281-287.
- [16] DENG Y,LI X Y,CUI J,et al. Multi head attentional memory network for short text emotion classification [J]. Computer Application,2021,41(11):3132-3138.
- [17] ZHANG X,LU W,ZHANG G,et al. Chinese sentence semantic matching based on multi-granularity fusion model[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham:Springer,2020:246-257.
- [18] CHEN Q,ZHU X,LING Z,et al. Enhanced LSTM for natural language inference[J]. arXiv:1609.06038,2016.
- [19] WU S H,PENG D L,YUAN W W,et al. MGSC:A Short Text Semantic Matching Model with Multi-granularity Semantic Intersection [J]. Small and Microcomputer Systems,2019,40(6):1148-1152.
- [20] CHEN Y,QIU X Y. Multi-task text semantic matching method combined with self-supervised learning [J]. Journal of Peking University(Natural Science Edition),2022,58(1):83-90.
- [21] MA X Y,FAN Y X,GUO J F,et al. Research and Analysis on Generalization and Migration of Short Text Matching [J]. Computer Research and Development,2022,59(1):118-126.
- [22] XU W F,YANG Y,ZHANG C F. A deep text semantic matching model integrating entity context features [J]. Journal of Wuhan University(Science Edition),2020,66(5):483-494.
- [23] JUN Y S. jieba [EB/OL]. <https://github.com/fxsjy/jieba>.
- [24] CHO K,MERRIENBOER B V,GULCEHRE C,et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).2014:1724-1734.
- [25] LIU X,CHEN Q,DENG C,et al. Lcqm: A large-scale chinese question matching corpus[C]//Proceedings of the 27th International Conference on Computational Linguistics.2018:1952-1962.
- [26] CHEN Y. Convolutional neural network for sentence classification[D]. Waterloo:University of Waterloo,2015.
- [27] MUELLER J,THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2016.
- [28] WANG Z,HAMZA W,FLORIAN R. Bilateral multi-perspective matching for natural language sentences[J]. arXiv:1702.03814,2017.
- [29] YU B H,WANG J C. Research on Chinese Semantic Matching in Siamese Network [J]. Small and Microcomputer System,2021,42(2):231-234.



ZANG Jie, born in 1979, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include big data, optimization theory and application.



WANG Yan, born in 1978, Ph.D, professor, is a member of China Computer Federation. Her main research interests include big data, Internet of Things and blockchain technology.

(责任编辑:何杨)