Journal of Computer Applications
计算机应用,2019,39(6):1639 - 1645
ISSN 1001-9081
CODEN JYIIDU
2019-06-10
http://www.joca.cn

文章编号:1001-9081(2019)06-1639-07
DOI:10.11772/j.issn.1001-9081.2018102184

基于注意力和字嵌入的中文医疗问答匹配方法

陈志豪[1*],余　翔[1],刘子辰[2],邱大伟[2],顾本刚[1]
(1.重庆邮电大学 通信与信息工程学院,重庆 400065;
2.移动计算与新型终端北京重点实验室(中国科学院 计算技术研究所),北京 100190)
(*通信作者电子邮箱 chenzhihao@ict.ac.cn)

摘　要:针对当前的分词工具在中文医疗领域无法有效切分出所有医学术语,且特征工程需消耗大量人力成本的问题,提出了一种基于注意力机制和字嵌入的多尺度卷积神经网络建模方法。该方法使用字嵌入结合多尺度卷积神经网络用以提取问题句子和答案句子不同尺度的上下文信息,并引入注意力机制来强调问题和答案句子之间的相互影响,该方法能有效学习问题句子和正确答案句子之间的语义关系。由于中文医疗领域问答匹配任务没有标准的评测数据集,因此使用公开可用的中文医疗问答数据集(cMedQA)进行评测,实验结果表明该方法优于词匹配、字匹配和双向长短时记忆神经网络(BiLSTM)建模方法,并且Top-1准确率为65.43%。

关键词:自然语言处理;问答对匹配;卷积神经网络;字嵌入;注意力机制
中图分类号:TP183　　**文献标志码**:A

Chinese medical question answer matching method based on attention mechanism and character embedding

CHEN Zhihao[1*], YU Xiang[1], LIU Zichen[2], QIU Dawei[2], GU Bengang[1]
(1. *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;
2. *Beijing Key Laboratory of Mobile Computing and Pervasive Device*
(*Institute of Computing Technology, Chinese Academy of Sciences*), *Beijing* 100190, *China*)

Abstract: Aiming at the problems that the current word segmentation tool can not effectively distinguish all medical terms in Chinese medical field, and feature engineering has high labor cost, a multi-scale Convolutional Neural Network (CNN) modeling method based on attention mechanism and character embedding was proposed. In the proposed method, character embedding was combined with multi-scale CNN to extract context information at different scales of question and answer sentences, and attention mechanism was introduced to emphasize the interaction between question sentences and answer sentences, meanwhile the semantic relationship between the question sentence and the correct answer sentence was able to be effectively learned. Since the question and answer matching task in Chinese medical field does not have a standard evaluation dataset, the proposed method was evaluated using the publicly available Chinese Medical Question and Answer dataset (cMedQA). The experimental results show that the proposed method is superior to word matching, character matching and Bi-directional Long Short-Term Memory network (BiLSTM) modeling method, and the Top-1 accuracy is 65.43 %.

Key words: natural language processing; question answer matching; Convolutional Neural Network (CNN); character embedding; attention mechanism

0　引言

随着互联网的快速发展,愈来愈多的人倾向于在健康医疗网站上提问来寻求健康帮助,例如中国的寻医问药网、39健康网和丁香园等。此类网站为患者和医生提供了一个在线交流的平台,便于用户随时随地获取高质量的医疗健康推荐。患者只需描述其自身的症状并发布问题,就能得到指定的医生或任意医生的回复和建议。然而,大多数情况下,许多用户提出的问题都相似,这一方面给医生专家带来了巨大的回复负担,另一方面延长了患者等待回复的时间。因此,为了提高用户体验,有必要设计一种方法来有效地处理医疗问答匹配的问题,即从已有的医疗答复记录中自动选择与用户问题匹配最佳的答复推荐给用户。

本文重点关注的是中文医疗问答匹配和答案选择的问题,其中所考虑的问答语言均限于中文。相比于Feng等[1]和Tan等[2]在英语语言环境下的开放领域问答匹配的研究,本文所讨论的问题更具挑战性,原因有两点:1)领域受限性质;2)中文语言具有一些特殊的特征。

进一步的讨论如下:

首先,由于汉语是以字为基本的书写单位,词语之间没有

收稿日期:2018-10-31;**修回日期**:2018-12-31;**录用日期**:2019-01-10。　**基金项目**:国家重大科技专项(2016ZX03002010-003)。
作者简介:陈志豪(1994—),男,重庆人,硕士研究生,主要研究方向:自然语言处理、智能问答系统;　余翔(1964—),男,四川成都人,正高级工程师,主要研究方向:数字通信、无线信号处理;　刘子辰(1985—),男,山东临沂人,助理研究员,主要研究方向:网络通信、大数据挖掘;邱大伟(1991—),男,内蒙古赤峰市人,博士研究生,主要研究方向:模式识别、机器学习、自然语言处理;　顾本刚(1992—),男,安徽淮南人,硕士研究生,主要研究方向:网络通信。

明显的区分标记，因此分词是大多数中文自然语言处理(Natural Language Processing，NLP)任务中不可或缺的数据预处理步骤，如词性标注(Part-of-Speech tagging，POS)和语义分析。由此可见，分词的准确与否大大影响了下游任务的准确性。尽管已有的分词工具(如：ICTCLAS、jieba 和 HanLP)的性能已经达到了满足大多数实际应用的水平，但它们一旦发生偏差，将通过管道不可避免的影响整个系统框架，导致整体性能下降[3]。此外，当直接应用于医学文本时，包含的各种医学术语会导致这些通用分词工具的性能进一步下降。例如，药物名称“活血止痛片”和“维生素 C 黄连上清片”被 jieba 分词工具错误的划分为了“活血 止痛 片”和“维生素 C 黄连 上清 片”。尽管引入特定领域的词典可以减轻专业术语对分词的负面影响，但构建此类词典似乎总是令人望而却步，因为它涉及大量的手工劳动并需要大量特定领域的专业知识。更糟糕的是，在处理在线社区发布的未经编辑的问题和答案时，预定义的词典往往不合适。因为它们通常是以非正式的表达形式编写的，往往包含许多简写词和非标准的缩略词，甚至是错误的拼写和不合适的语法结构的句子。例如，问句“嘴骑魔拖车摔肿了怎么消下去现在还疼不能吃饭感觉越来越大也不能喝水”将“摩托车”误写成了“魔拖车”，另外全句没有任何标点符号，逻辑表达混乱。虽然，通用分词工具都能够加载定制的领域词典，但是定制词典需要耗费大量的人工时间，且定制的词典也不可能覆盖所有的领域词。

为了避免上述问题，提出采用字嵌入的端到端的神经网络框架。该框架采用的是字级的表示，即用字嵌入方式替代传统的词嵌入方式。此种方式既可以避免数据预处理时的分词步骤，也可以避免由分词错误引起的其他组件的性能下降。问题和答案的表示向量分别使用中文字进行预训练得到，并且类似于词嵌入，将每个字描述为固定长度的向量。

由于在中文语言中字所含的语义信息比词语的语义信息少，若采用统计方法则可能需要使用语言模型或词性标注等方式来抽取相关的语义信息。然而，卷积神经网络(Convolutional Neural Network，CNN)强调 *N*-Gram 内的本地交互，能够自动捕获字和词语的局部语义信息，无需其他方法辅助，因此本文引入 CNN 来构建模型。又因为中文词语或短语通常由 2 至 5 个字构成，所以采用多尺度卷积神经网络(Multi-scale CNNs，MultiCNNs)来提取不同尺度的上下文信息，由此可以更好地编码问题和答案。因此，本文提到的 MultiCNNs 模型由一组不同尺度的卷积核组成。

大多数之前的工作都是将问题和答案两个句子分别表示，很少考虑一个句子对另一个句子的影响。这忽略了两个句子在同一任务背景下的相互影响，也与人类在比较两个句子时的行为相矛盾。人们通常从另一个句子中提取与身份、同义词、反义词和其他关系相关的部分来找到一句话中的关键部分。受 Yin 等[4]提出的基于注意力(Attention)机制对句子对联合建模的方法的启发，本文引入 Attention 机制将问题和答案两个句子一起建模，用一个句子的内容来指导另一个句子的表示。因此，本文提出了基于注意力和字嵌入的卷积神经网络(CNNs based on Attention Mechanism and character embedding，AMCNNs)框架。

答案推荐是问答系统的目标。对于每个问题描述，答案候选池都包含 100 个候选答案，其中有一个或多个相关答案和多个不相关答案。问答匹配任务的目标就是从候选答案中找出精确度分数最高(Top-1)的一个答案，并将其推荐给用户。

1 国内外研究现状

1.1 传统问答方法

Jain 等[5]提出了基于规则的医学领域问答系统架构，并详细讨论了基于规则的问题处理和答案检索的复杂性。然而，由于用户问题总是以大量不同的方式呈现，因此基于规则的方法可能无法涵盖所有表达方式。

Wang 等[6]提出了另一种方法，首先将句子划分成单词以训练每个句子的词向量，然后通过计算每个单词之间的相似性来评估每个问答对的相似性。而 Abacha 等[7]则是将问题翻译成机器可读的表示形式。因此，该方法能够将各种自然语言表达问题转换为标准的表示形式。后来，Abacha 等[8]通过在表示和查询层应用语义技术来扩展他们以前的工作，以便创建结构化查询以匹配知识库中的条目。这些方法取决于手工设计的模式和特征，需要巨大的人力和专业知识。

现有研究提出了一些关于中文问答的模型。Li 等[9]构建了一个用于音乐领域的语义匹配模型，能够自动将问题翻译成 SPARQL 查询语句来获得最终答案。Yin 等[10]针对在线健康专家问答服务效率低下的问题，开发了用于对相似的问题和答案进行分组的分层聚类方法和用于检索相关答案的扩展相似性评估算法，用于从已有的专家答案中进一步提取出答案。然而，这些方法都将分词作为中文文本处理的必要步骤，虽然在一般领域中分词工具能达到研究者们的期望，但它们并未考虑误差所带来的影响。Wang 等[11]提出了一种集成基于计数和基于嵌入的特征的方法；他们还在研究中指出，基于字的模型优于基于词的模型，从中得到启示：处理汉字可以避免分词错误带来的不利影响。

1.2 深度学习在问答匹配任务中的应用

近年来，由于深度学习技术无需任何语言工具、特征工程或其他资源，因此，愈来愈多的自然语言处理(NLP)任务都采用了该技术。

Feng 等[1]设计了 6 个深度学习模型，并在保险领域进行了问答匹配的实验，该实验结果为其他问答匹配任务研究者提供了有价值的指导(例如，卷积层后不需要全连接层)。Hu 等[12]提出了两种不同的卷积模型来学习句子的表示，这是使用神经网络解决一般句子匹配问题的先驱工作。之后，Feng 等[1]和 Zhou 等[13]采用 CNNs 来学习问答对的表示，进一步用于计算不同问题与候选答案之间的相似性。后来，为了从句子中提取序列信息，Tan 等[2,14]利用递归神经网络(Recurrent Neural Network，RNN)及其变体长短期记忆网络(Long Short-Term Memory network，LSTM)来学习句子级表示；值得注意的是，他们还利用注意力机制来增强问题和答案之间的语义关联。Yin 等[4]设计了 3 种基于注意力机制的卷积神经网络(Attention-Based CNN，ABCNN)来建模问答对，并分别在答案选择(Answer Selection，AS)、释义识别

(Paraphrase Identification, PI)和文本蕴含(Textual Entailment, TE)3 个任务上进行了模型验证。Zhang 等[15]提出端到端的字嵌入多尺度卷积神经网络模型,用于医疗领域的问答匹配任务。

但是,上述所有研究都与英文文本相关。当直接用于处理中文文档时,所提出方法的性能会出现相当大的下降,是由于中文与英文的结构有很大不同。

2　AMCNNs

本文将详细介绍用于中文医疗问答对匹配的多尺度卷积神经网络模型,其基于注意力机制和字嵌入。首先,本文讨论编码中文医疗文本的正确嵌入方式;其次,详细描述字嵌入多尺度卷积神经网络架构和本文提出的基于注意力机制的字嵌入多尺度神经网络架构。

2.1　字符的分布式表示

在许多自然语言处理任务中,一个基本的步骤就是将文本序列转换成机器可读的特征,该特征通常是固定长度的向量。

近年来,基于嵌入的方式在文本特征提取中得到了广泛应用,证明了它在语义表示上的效用。而用得最多的一种嵌入方式就是词嵌入(word-embedding)方式,也即分布式词表示。Bengio 等[16]提出了一个神经网络语言模型(Neural Network Language Model, NNLM),它将神经网络与自然语言处理相结合以训练词嵌入。之后,Mikolov 等[17]受 NNLM 的启发提出了一个非常高效的语言模型:Word2Vec。而且,近年来 Word2Vec 受到越来越多的关注并成功应用于许多 NLP 任务,例如句子匹配[18]、文档分类[19]和知识图谱抽取[20]。

Word2Vect 模型的一个例子如图 1 所示。给定句子序列 $Sent = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_l]$,其中 l 是序列的长度。$\boldsymbol{w}_i \in \mathbf{R}^{d_w}$ 表示句子中位置为 i 的词的词向量。使 $Context(\boldsymbol{w}_i) = [\boldsymbol{w}_{i-k}, \boldsymbol{w}_{i-k+1}, \cdots, \boldsymbol{w}_{i-1}, \boldsymbol{w}_{i+1}, \cdots, \boldsymbol{w}_{i+k-1}, \boldsymbol{w}_{i+k}]$ 表示词 $\boldsymbol{w}_i$ 的上下文,其中 $2k$ 表示上下文窗口的尺寸。而使 $p(\boldsymbol{w}_i \mid Context(\boldsymbol{w}_i))$ 表示句子中第 i 个词是 $\boldsymbol{w}_i$ 的概率。该模型的目标是优化对数最大似然估计($\log(MLE)$):

$$L_w(MLE) = \sum_{\boldsymbol{w}_i \in S} \log(p(\boldsymbol{w}_i \mid Context(\boldsymbol{w}_i))) \tag{1}$$

然而,当遇到未登录词或稀缺词时,word-embedding 方式在中文处理上的质量会有所下降。

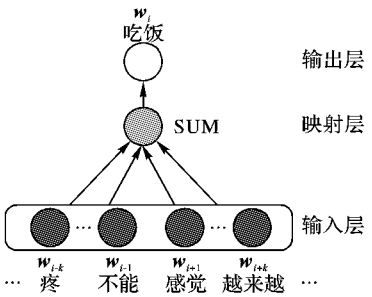

图 1　词的分布式表示

Fig. 1　Distributed representation of words

受 Wang 等[11]的工作的启发,本文将句子分成单独的字,如图 2 所示,在上下文窗口中,每个字被用来预测它们中间的字。利用 gensim 工具训练字向量,训练完之后,每个字就被映射到了一个固定长度的向量 $\boldsymbol{c}_i \in \mathbf{R}^{d_c}$,其中 d_c 表示向量的维度。

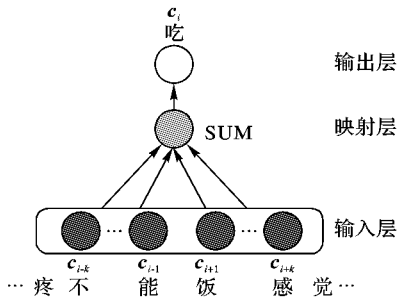

图 2　字的分布式表示

Fig. 2　Distributed representation of characters

给定句子序列 $Sent = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_l]$,其中 l 表示序列中字的数量,而 $\boldsymbol{c}_i$ 的上下文表示则是 $Context(\boldsymbol{c}_i) = [\boldsymbol{c}_{i-k}, \boldsymbol{c}_{i-k+1}, \cdots, \boldsymbol{c}_{i-1}, \boldsymbol{c}_{i+1}, \cdots, \boldsymbol{c}_{i+k-1}, \boldsymbol{c}_{i+k}]$。因此,式(1)可以修改为:

$$L_c(MLE) = \sum_{\boldsymbol{c}_i \in S} \log(p(\boldsymbol{c}_i \mid Context(\boldsymbol{c}_i))) \tag{2}$$

字嵌入能够避免分词算法引发的错误造成的影响[21]。此外,由于字的数量比词的数量少,新字或稀有字的数量也比新词或稀有词的数量少,因此字级的表示方式能降低句子的表示向量尺寸。目前,字嵌入方式已经应用于许多自然语言处理任务,例如机器翻译[22]、文本分类[23]、英文场景问答[24]和汉语句法依赖解析[21]。但是与英文相比,中文字嵌入方式主要是为了缓解分词不准确带来的影响。另外,汉字的数量比英文字母数量(只有 26 个)多太多且中文文本中的汉字也比英文的字符包含更多的信息。

2.2　多尺度卷积神经网络架构

目前,卷积神经网络因其能够捕获本地上下文信息的能力而在许多 NLP 任务中得到了应用。该网络不依赖于其他如词性标注或解析树之类的外部信息。通常,一个卷积神经网络架构由两部分组成:卷积和池化。卷积步骤利用固定大小的滑动窗口提取本地上下文特征,而池化步骤选择从前一层提取的特征的最大值或平均值以降低输出维度,但保留了最突出的信息。然而,单尺度卷积神经网络(Single-scale CNN, SingleCNN)架构只有一个固定卷积窗口的特征映射,也即它能捕获的信息很少。考虑到汉语短语的表达结构,SingleCNN 架构可能不足以提取字信息。而多尺度的卷积神经网络(MultiCNNs)架构的工作方式与 SingleCNN 架构的工作方式相似,唯一不同之处在于采用了多个不同尺度的特征映射来提取信息。该架构如图 3 所示,问题和答案句子分别由固定长度的字符嵌入序列表示:$[\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_l]$。字向量的维度由 d_c 表示,且向量中的每个元素都是实数,则 $\boldsymbol{c}_i \in \mathbf{R}^{d_c}$。每个句子需要归一化为一个固定长度的序列,即若句子长度小于某个阈值就添加 0 补齐,相反若大于某个阈值就裁剪掉多余部分。经过字嵌入层后,每个问题和答案分别由矩阵 $\boldsymbol{Q}_e \in \mathbf{R}^{l \times d_c}$ 和 $\boldsymbol{A}_e \in \mathbf{R}^{l \times d_c}$ 表示。

假设存在一个卷积核尺寸集合 $S = \{s_1, s_2, \cdots, s_t\}$,其中第 i 个卷积神经网络卷积核的尺寸表示为 s_i。当给定序列 $\boldsymbol{Z} = [z_1, z_2, \cdots, z_{l-s_i+1}]$,其中 $z_i = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_{i+s_i-1}]$ 表示句子中连

续 s_i 个字向量拼接的结果，也即每个特征映射提取出的信息。因此，可以定义卷积运算的计算式如下：

$$\boldsymbol{O}_j^{s_i} = f(\boldsymbol{w}_j^{s_i} \cdot [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_{l-s_i+1}] + \boldsymbol{b}^{s_i}) \tag{3}$$

式中：$\boldsymbol{O}_j^{s_i} \in \mathbf{R}^{l-s_i+1}$；矩阵 $\boldsymbol{w}_j^{s_i} \in \mathbf{R}^{s_i \times d_c}$ 和 $\boldsymbol{b}^{s_i}$ 是需要训练的参数；$f(\cdot)$ 表示激活函数；$\boldsymbol{W} \cdot \boldsymbol{Z}$ 表示矩阵 $\boldsymbol{W}$ 和矩阵 $\boldsymbol{Z}$ 的对应元素相乘。若特征映射的数量为 d，则卷积层的输出为 $\boldsymbol{O}^{s_i} = [\boldsymbol{O}_1^{s_i}, \boldsymbol{O}_2^{s_i}, \cdots, \boldsymbol{O}_d^{s_i}]$。由此可见，嵌入矩阵 $\boldsymbol{Q}_e \in R^{l \times d_c}$ 和 $\boldsymbol{A}_e \in \mathbf{R}^{l \times d_c}$ 通过卷积神经层并共享相同卷积参数（$\boldsymbol{w}_j^{s_i}$ 和 $\boldsymbol{b}^{s_i}$）之后转变为矩阵 $\boldsymbol{Q}_o \in \mathbf{R}^{(l-s_i+1) \times d}$ 和 $\boldsymbol{A}_o \in \mathbf{R}^{(l-s_i+1) \times d}$。

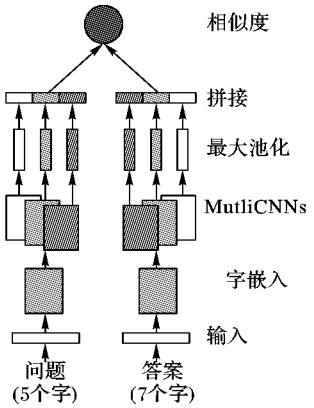

图 3 多尺度卷积神经网络架构

Fig. 3 Framework of multi-scale CNNs

在卷积层之后应用池化层来抽取出最突出的信息，可以选择最大池化和平均池化，本文选择使用最大池化方式从每个过滤器中选择出最大值。经过池化层后，第 i 个卷积层的输出就可以表示为：

$$\boldsymbol{p}^{s_i} = [\max \boldsymbol{O}_1^{s_i}, \max \boldsymbol{O}_2^{s_i}, \cdots, \max \boldsymbol{O}_d^{s_i}] \tag{4}$$

式中：$\max \boldsymbol{O}_j^{s_i} (j \in [1,d])$ 表示从 $\boldsymbol{O}_j^{s_i}$ 中选择最大的值，由此 $\boldsymbol{p}^{s_i} \in \mathbf{R}^d$。

与 SingleCNN 架构不同的是，在经过最大池化层之后，需要分别将不同尺度过滤器的最大池化结果向量拼接为一个长向量作为问题或答案的最终表示：

$$\boldsymbol{p} = [\boldsymbol{p}^{s_1}, \boldsymbol{p}^{s_2}, \cdots, \boldsymbol{p}^{s_t}] \tag{5}$$

本文用向量 $\boldsymbol{Q}_p \in \mathbf{R}^{t \times d}$ 和 $\boldsymbol{A}_p \in \mathbf{R}^{t \times d}$ 分别表示问题和答案经过卷积、最大池化和拼接之后的最终表示形式，然后通过式（6）计算它们的相似度：

$$Sim(\boldsymbol{Q}_p, \boldsymbol{A}_p) = Cos(\boldsymbol{Q}_p, \boldsymbol{A}_p) = \frac{\| \boldsymbol{Q}_p \cdot \boldsymbol{A}_p \|}{\| \boldsymbol{Q}_p \| \times \| \boldsymbol{A}_p \|} \tag{6}$$

式中：$\| \cdot \|$ 表示向量的长度。

多尺度卷积神经网络架构完全能够从中文文本中提取相关的语言特征。图 4 充分展示了使用 MultiCNNs 架构从一个固定长度区域句子中提取局部上下文信息的过程。已经知道，不同的中文短语通常包含不同数量的字。因此，单尺度卷积神经网络将在固定长度区域上执行卷积运算，这类似于将几个相邻字组合成单词。因此，MultiCNNs 架构在不同的固定长度区域上执行卷积运算，并提取不同数量的相邻字嵌入。

2.3 基于注意力机制的多尺度卷积神经网络架构

基于注意力机制的多尺度卷积神经网络（AMCNNs）架构的工作方式与 MultiCNNs 相似，唯一不同之处在于两个句子经过嵌入层后会分别与注意力特征矩阵 $\boldsymbol{A}$ 相乘，得到一个注意力特征映射矩阵，并与经过嵌入层后得到的矩阵一起作为卷积层的输入。该架构如图 5 所示。由于考虑到两个句子在同一任务背景下的相互影响，因此用一个句子的内容来指导另一个句子的表示。

AMCNNs 利用一个注意力特征矩阵 $\boldsymbol{A}$ 来影响卷积运算。注意力特征旨在对卷积中与问题（或答案）单位相关的答案（或问题）单位赋予更高的权重。如图5所示，矩阵 $\boldsymbol{A}$ 是通过左侧的字嵌入表示单元和右侧的字嵌入表示单元相匹配产生的。$\boldsymbol{A}$ 中第 i 行的注意值表示 s_q 的第 i 个单位相对于 s_a 的注意力分布，$\boldsymbol{A}$ 中第 j 列的注意值表示 s_a 的第 j 个单位相对于 s_q 的注意力分布。$\boldsymbol{A}$ 可以视为 s_q（或 s_a）在行（或列）方向的新的嵌入表示，因为 $\boldsymbol{A}$ 的每一行（或列）是 s_q（或 s_a）上的一个字的新的特征向量。因此，将这个新的特征向量与字嵌入表示组合并将它们用作卷积运算的输入就有理可依了。通过将注意力特征矩阵 $\boldsymbol{A}$ 转换为图 5 中的两个与字嵌入表示相同格式的灰色矩阵来实现这一点。因此，卷积层的新输入具有每个句子的两个特征映射。

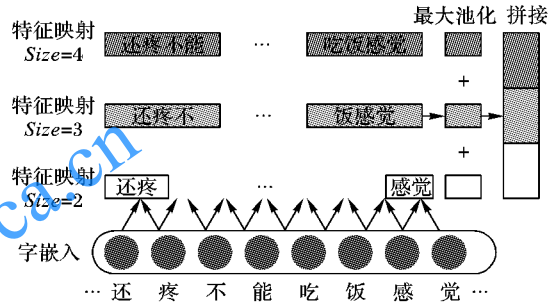

图 4 MultiCNNs 架构提取上下文信息的过程

Fig. 4 Process of extracting context information by MultiCNNs framework

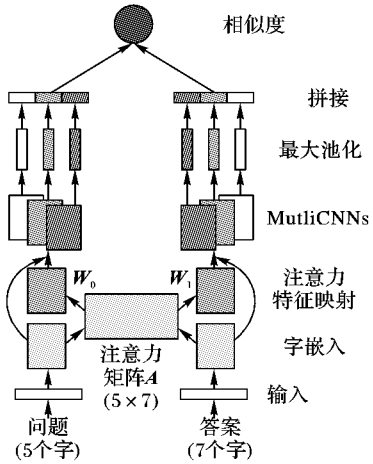

图 5 基于注意力机制的多尺度卷积神经网络（AMCNNs）架构

Fig. 5 Framework of AMCNNs

接下来，采用文献[4]的方法定义注意力矩阵 $\boldsymbol{A} \in \mathbf{R}^{d_c \times d_c}$，如式（7）所示：

$$\boldsymbol{A}_{i,j} = match\text{-}score(\boldsymbol{Q}_e[i,:], \boldsymbol{A}_e[j,:]) \tag{7}$$

式中 $match\text{-}score(\cdot)$ 函数有多种方式定义，但本文采用的是欧几里得距离算法定义该函数，即 $\frac{1}{1+(x-y)}$。

给定注意力矩阵 $\boldsymbol{A}$，以此分别产生问题和答案的注意力映射矩阵，计算式如下：

$$\begin{cases} \boldsymbol{F}_{Q,a} = \boldsymbol{W}_0 \cdot \boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{F}_{A,a} = \boldsymbol{W}_1 \cdot \boldsymbol{A} \end{cases} \tag{8}$$

式中:权重矩阵 $\boldsymbol{W}_0 \in \mathbf{R}^{l \times d_c}$, $\boldsymbol{W}_1 \in \mathbf{R}^{l \times d_c}$ 都是模型训练时需要学习的参数。

分别得到问题和答案的注意力映射矩阵之后,然后将它们组合为 3 阶的张量传递给卷积层进行卷积运算。卷积层的结构和运算过程以及在此之后的池化步骤都与图 4 的 MultiCNNs 架构一样。最后将最大池化后的问题向量和答案向量进行相似度计算,从而得到它们的匹配结果。

2.4　目标函数

给定一个问题 q_i,它的真实答案是 a_i^+,而 a_i^- 是从候选答案池中随机选择的错误答案。一个有效的网络模型应该能够最大化 $Sim(q_i, a_i^+)$,且最小化 $Sim(q_i, a_i^-)$。为了训练以上神经网络,本文采用文献[1,4,13,16 - 17]中的方法定义训练损失函数为:

$$H = \max\{0, M - Sim(q_i, a_i^+) + Sim(q_i, a_i^-)\} \tag{9}$$

式中:M 是一个常数,表示边界值。如果 $Sim(q_i, a_i^+) - Sim(q_i, a_i^-) > M$,则损失函数为 0,且网络参数不需要更新。

当训练该网络时,分别提供每个问题与其真实答案和随机采样的错误答案给网络。然后,计算损失函数 H,并且使用优化器(如 GradientDescentOptimizer 或 AdagradOptimizer)以更新网络参数。

3　数据集和实验设置

3.1　cMedQA 数据集

当前在中文医疗领域问答匹配任务上没有标准的评测数据集,因此选择从专业的医疗健康网站上收集中文医疗问答数据集(Chinese Medical Question and Answer dataset, cMedQA)[15]。该数据集是一个基于中文医疗问答的公开可用的数据集,其中的问答数据收集于寻医问药网。问题通常是由用户提出,而答案来自于专业医生的可靠回复。通常一个问题会得到多个医生的回复,即一个问题有多个正确答案。

如表 1 所示,该数据集已经被划分为了三个子集:训练集、开发集和测试集。

表 1　cMedQA 数据集的统计信息

Tab. 1　Statistical information for cMedQA dataset

数据集	问题句子数	答案句子数	每句问题平均词数	每句答案平均词数	每句问题平均字数	每句答案平均字数
训练集	50 000	94 134	97	169	120	212
开发集	2 000	3 774	94	172	117	216
测试集	2 000	3 835	96	168	119	211
总计	54 000	101 743	96	169	119	212

训练该模型时,训练集中的每个问题 q_i 都有一个正确答案 a_i^+ 和一个从候选答案池中随机采样的答案 a_i^-,从而设定训练集中每个问题关联 30 个 (q_i, a_i^+, a_i^-) 问题答案对。因此,在训练阶段,每一轮总共有 1 500 000 问答对被依次送入网络中进行训练。开发集和测试集则是每个问题对应 100 个候选答案,其中包括一个或多个正确答案。开发集用于网络参数的优化,而测试集用于评估模型。

3.2　评价方法

Top-K 精度(ACC@K)普遍作为信息检索任务的评价标准,其定义如式(10)所示:

$$ACC@K = \frac{1}{N}\sum_{i=1}^{N} 1[a_i \in c_i^k] \tag{10}$$

式中:a_i 代表问题 q_i 的正确答案;c_i^k 指候选池中相似度分数最高的 K 个答案。$a_i \in c_i^k$ 是函数 $1[\cdot]$ 的条件,当条件满足时,其值为 1;否则,其值为 0。

由于答案选择(AS)任务与信息检索任务不同,AS 要求得到一个可能性最大的答案,即返回候选答案池相似度分数最高的一个答案。因此,本文选择 top-1 精度(ACC@1)作为本文模型评估的标准。

3.3　基线模型

目前优秀的问答匹配模型都是基于英文语境,且面向开放领域问答,因此本文所描述模型不与它们进行比较,而是设计了一些基线模型用于实验效果对比,如下所示:

1)字匹配(Character Matching)。字匹配方式统计问题与答案中相同字的数量。

2)词匹配(Word Matching)。与字匹配相似,统计问题与答案中相同词语的数量。但是需要使用分词工具进行分词,因此本文选择了两种分词工具(Jieba 和 ICTCLAS)用以展示不同分词工具对模型性能的影响。

3)BM25。BM25(最佳匹配)是信息检索中的一种排序函数,该函数定义如下:

$$BM25(q_i, a_i) = \sum_{w_j \in q_i} \frac{IDF(w_j) \cdot f(w_j, a_i) \cdot (k+1)}{f(w_j, a_i) + k \cdot (1 - b + b \cdot |a_i| / |a|_{\mathrm{avg}})} \tag{11}$$

式中:$IDF(w_j)$ 是问句中词(或字)w_j 的逆文档频率;$|a|$ 是答案 a_i 的长度,$|a|_{\mathrm{avg}}$ 是答案的平均长度;$f(w_j, a_i)$ 表示在 a_i 中的 w_j 的频率。k 和 b 是需要设定的参数,在本文中,设定 $k = 2.0$, $b = 0.75$.

4)BiLSTM。Tan 等[2]使用双向长短时记忆神经网络(Bi-directional LSTM, BiLSTM)来学习问题和答案的语义表示。BiLSTM 是循环神经网络(RNN)的变体,能够捕获长文本序列的语义信息。

3.4　实验设置

本文提出的模型均是使用深度学习框架 TensorFlow 训练和测试的。为了比较字符嵌入和词嵌入的性能,选择使用 Jieba 和 ICTCLAS 两个分词工具用于分词。同时本文使用 gensim 工具预先训练字向量和词向量,训练完之后的每个字(词)由 $d_c = 200$ 的一维向量表示。字向量在模型训练过程中也是作为参数被优化。在模型训练中,批次数量设置为 100,而问题和答案的最大序列长度均为 400 字符和 200 词。

对于 CNN 网络,滤波器具有 3 种不同大小的尺寸,分别为 2、3、4,但特征映射数量均为 800 个。在本文中使用 AdagradOptimizer 优化器,学习率初始化值设置为 0.01,边界值 M 设置为 0.5。

3.5　结果分析

在 cMedQA 数据集[15]上采用不同方法得到的实验结果如表 2 所示。

如表 2 所示,第 1 ~6 行提供的是没有使用神经网络架构的基线模型的实验结果。其中,第 1 ~3 行展示的是基于匹配的方式得到的结果,可以发现词匹配和字匹配的效果相差 1 个百分点左右,这是由于词中包含了比字更多的信息。相较于词(字)匹配的方式,BM25 方式利用了更多的统计信息得到了 11 个百分点的提升。

第 7 ~9 行展示的是基于嵌入的方式得到的结果。从中可以发现,字嵌入的效果明显由于词嵌入,说明能够从字中更好地提取嵌入方式的语义信息;而使用 ICTCLAS 分词工具得到的词嵌入结果优于使用 jieba 分词工具得到的结果,表明了分词工具的性能对实验结果有较大的影响。

第 10 ~12 行展示了词嵌入的深度神经网络进行建模的结果,其使用 jieba 分词工具进行文本预处理。从中可以发现,BiLSTM[2] 网络架构的效果显著优于卷积神经网络架构。原因可能是循环神经网络及其变体能够抽取整个句子的语义信息,能有效降低问题和答案的语义代沟。该实验结果与 Zhang 等[15] 的实验结果一致。

表 2 各模型的 Top-1 准确率结果

Tab. 2 Top-1 accuracy results of different model

编号	嵌入方式	模型	准确率/%	
			开发集	测试集
1	None	Word Matching(Jieba)	37.05	36.60
2		Word Matching(ICTCLAS)	35.11	36.22
3		Character Matching	33.65	34.90
4		BM25(Jieba)	37.60	40.00
5		BM25(ICTCLAS)	40.25	41.25
6		BM25(character)	44.80	45.40
7	Word(Jieba)	Embedding Matching	24.55	23.65
8	Word(ICTCLAS)		27.85	29.10
9	Character		30.80	32.30
10	Word Embedding (Jieba)	BiLSTM	51.70	50.10
11		MultiCNNs	48.40	47.75
12		AMCNNs	51.25	50.01
13	Word Embedding (ICTCLAS)	BiLSTM	56.15	56.02
14		MultiCNNs	53.06	52.34
15		AMCNNs	53.50	51.78
16	Character Embedding	BiLSTM	61.65	60.78
17		MultiCNNs	65.35	64.73
18		AMCNNs	66.80	65.43

第 13 ~15 行也是采用了词嵌入的深度神经网络建模的结果,其使用 ICTCLAS 分词工具进行文本预处理。与第 10 ~12 行的实验对比结果一样,BiLSTM 的效果仍然优于卷积神经网络。同时,可以发现使用 ICTCLAS 分词工具进行文本预处理后进行建模的方法优于使用 Jieba 分词工具的方法。并且在词嵌入方式中比较了未加 Attention 机制(MultiCNNs)和加了 Attention 机制(AMCNNs)的多尺度卷积神经网络的性能。通过对比可以看出,不管是采用 Jieba 工具还是 ICTCLAS 工具,都是结合注意力机制的卷积神经网络架构优于卷积神经网络架构,同时进一步证明了分词工具对模型的性能有较大的影响。因此,为了避免分词工具的影响,采用字嵌入方式是有必要的。

第 16 ~ 18 行采用字嵌入方式分别训练了 BiLSTM、MutliCNNs 和 AMCNNs。从结果中可以发现,MultiCNNs 的准确率比 BiLSTM 提升了约 4 个百分点,而 AMCNNs 的准确率相较于 MultiCNNs 则提升了约 2 个百分点,表明增加的 Attention 机制是能够改善 MutliCNNs 架构的性能的。同时,对比第 10 ~15 行的效果能够看出,字嵌入方式明显优于词嵌入,准确率提升了约 10 个百分点。

由以上实验结果可得,基于神经网络的方式明显优于词(字)匹配方式,说明了语言表达的多样性和非正规性很有可能降低匹配的效果;而字嵌入方式明显优于词嵌入,不论是单独使用还是在神经网络中。以上结论表明本文提出的基于注意力机制的字嵌入多尺度卷积神经网络用于医疗领域问答匹配是可行的。

通过对 CNN 的研究可知,该网络架构处理文本的原理类似于 N-Gram 语言模型,N 一般在[2,4]。同时,考虑到中文词语通常由 2 至 4 个汉字构成,即卷积操作的窗口尺寸大小应该在[2,4]。因此本文对比了 2,3,4 三种尺度的不同组合的效果,结果如表 4 所示。根据实验结果,最终选择(2,3,4)的尺度组合。从表 4 中可以明显看出,当设置卷积核尺度为(3,4)时得到的效果最优。究其原因可能因为语料库中由 3 个字或 4 个字组合成的短语包含的信息更加丰富。

表 3 不同尺度卷积核的实验结果

Tab. 3 Experimental results of convolution kernels at different scales

卷积核尺度	准确率/%		卷积核尺度	准确率/%	
	开发集	测试集		开发集	测试集
(3, 4)	64.90	64.33	(2, 4)	64.95	64.15
(2, 3)	65.00	64.30	(2, 3, 4)	65.80	64.43

4 结语

本文设计并实现了基于注意力机制的字嵌入多尺度卷积神经网络模型,该模型用于处理中文医疗领域的问答匹配问题。同时,该模型不需要任何额外的特征工程、句法信息或者基于规则的模板。根据实验结果可知,该模型优于词嵌入方式,而且相较于无注意力机制的多尺度卷积神经网络和 BiLSTM,能更好地捕获字符级信息。

在接下来的工作中,我们将研究新的注意力机制,以便能更加准确地找到问题和答案之间的语义联系,以此来进一步提升该模型的效果。同时,我们将扩展本文模型,使其能够有效融合医疗领域的知识库并应用于智能问答系统中。

参考文献 (References)

[1] FENG M W, XIANG B, GLASS M R, et al. Applying deep learning to answer selection: a study and an open task [C] // Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway, NJ: IEEE, 2015: 813 - 820.

[2] TAN M, dos SANTOS C N, XIANG B, et al. Improved representation learning for question answer matching [C] // ACL 2016: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Cambridge, CA: MIT Press, 2016: 464 - 473.

[3] QIU X P, HUANG X J. Convolutional neural tensor network architecture for community-based question answering [C] // IJCAI 2015:

Proceedings of the 24th International Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2015: 1305 – 1311.

[4] YIN W P, SCHÜTZE H, XIANG B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs [EB/OL]. [2018- 08- 20]. http://cn. arxiv. org/abs/1512. 05193. pdf.

[5] JAIN S, DODIYA T. Rule based architecture for medical question answering system [C] // SocProS 2012: Proceedings of the Second International Conference on Soft Computing for Problem Solving, AISC 236. Berlin: Springer, 2014: 1225 – 1233.

[6] WANG J, MAN C T, ZHAO Y F, et al. An answer recommendation algorithm for medical community question answering systems [C] // SOLI 2016: Proceedings of the 2016 IEEE International Conference on Service Operations and Logistics, and Informatics. Piscataway, NJ: IEEE, 2016: 139 – 144.

[7] ABACHA A B, ZWEIGENBAUM P. Medical question answering: translating medical questions into SPARQL queries [C] // Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. New York: ACM, 2012: 41 – 50.

[8] ABACHA A B, ZWEIGENBAUM P. MEANS: a medical question-answering system combining NLP techniques and semantic Web technologies [J]. Information Processing & Management, 2015, 51 (5): 570 – 594.

[9] LI T C, HAO Y, ZHU X Y, et al. A Chinese question answering system for specific domain [C] // WAIM 2014: Proceedings of the International Conference on Web-Age Information Management. Berlin: Springer, 2014: 590 – 601.

[10] YIN Y S, ZHANG Y, LIU X, et al. HealthQA: a Chinese QA summary system for smart health [C] // CSH 2014: Proceedings of the 2nd International Conference on Smart Health, LNCS 8549. Cham: Springer, 2014: 51 – 62.

[11] WANG B Y, NIU J B, MA L Q, et al. A Chinese question answering approach integrating count-based and embedding-based features [C] // Proceedings of the 2016 International Conference on Computer Processing of Oriental Languages, National CCF Conference on Natural Language Processing and Chinese Computing, LNCS 10102. Cham: Springer, 2016: 934 – 941.

[12] HU B T, LU Z D, LI H, et al. Convolutional neural network architectures for matching natural language sentences [C] // NIPS 2014: Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, CA: MIT Press, 2014: 2042 – 2050.

[13] ZHOU X Q, HU B T, CHEN Q C, et al. Answer sequence learning with neural networks for answer selection in community question answering [EB/OL]. [2018- 08- 14]. https://arxiv. org/abs/1506. 06490. pdf.

[14] TAN M, dos SANTOS C N, XIANG B, et al. LSTM-based deep learning models for non-factoid answer selection [EB/OL]. [2018-08-20]. https://arxiv. org/abs/1511. 04108. pdf.

[15] ZHANG S, ZHANG X, WANG H, et al. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs [J]. Applied Sciences, 2017, 7(8): 767.

[16] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137 – 1155.

[17] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2018-08-23]. http://papers. nips. cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality. pdf.

[18] WANG Z G, HAMZA W, FLORIAN R. Bilateral multi-perspective matching for natural language sentences [EB/OL]. [2018-08-20]. https://arxiv. org/abs/1702. 03814. pdf.

[19] TADDY M. Document classification by inversion of distributed language representations [EB/OL]. [2018-08-20]. https://arxiv. org/abs/1504. 07295. pdf.

[20] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion [C] // AAAI 2014: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 2181 – 2187.

[21] ZHANG M S, ZHANG Y, CHE W X, et al. Character-level Chinese dependency parsing [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Cambridge, MA: MIT Press, 2014: 1326 – 1336.

[22] CHUNG J, CHO K, BENGIO Y. A character-level decoder without explicit segmentation for neural machine translation [EB/OL]. [2018-08-15]. https://arxiv. org/abs/1603. 06147. pdf.

[23] ZHANG X, ZHAO J B, LECUN Y. Character-level convolutional networks for text classification [C] // NIPS 2015: Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 649 – 657.

[24] GOLUB D, HE X D. Character-level question answering with attention [C] // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2016: 1598 – 1607.

This work is partially supported by National Science and Technology Major Project (2016ZX03002010-003).

CHEN Zhihao, born in 1994, M. S. candidate. His research interests include natural language processing, intelligent question answer system.

YU Xiang, born in 1964, Ph. D., senior engineer. His reasearch interest includes digital communication, wireless signal processing.

LIU Zichen, born in 1985, Ph. D., research assistant. His research interests include telecommunication, big data mining.

QIU Dawei, born in 1991, Ph. D. candidate. His research interests include pattern recognition, machine learning, natural language processing.

GU Bengang, born in 1994, M. S. candidate. His research interest includes telecommunication.