**第四章 基于深度学习的答案选择排序模型**

4.1 基于CNN的句子对建模

答案选择排序模型是针对问题答案对进行建模，需要对问句和候选答案分别应用卷积神经网络，第二章中已经介绍过对于单个句子进行卷积，池化，到生成分布式语义表示的过程，本章在第二章的基础上，针对问题答案对分别建立卷积神经网络，假设滤波器的大小设置为k,对相邻k个词语做卷积之后进行池化操作，执行average-pooling取卷积结果的均值作为最终句子的分布式语义表示，具体计算过程如下：





其中，和是代表滤波器的权值矩阵，和是偏执向量，代表第个词语到第个词语的词向量，和分别代表问句的长度和候选答案的长度，即问句和候选答案中所包含的词的个数，将上述公式中的计算过程用图示的方法进行表示，可得到答案选择排序中基于CNN的句子对建模模型，如图4-2所示：



图5：答案选择排序中基于CNN的句子对建模模型

在前面的第三章中已经介绍了词向量计算的方法并生成了涵盖法律领域的词向量库，上述模型中，首先通过查找词向量库，将问句和候选答案中包含的词的词向量进行组合，以词向量矩阵的形式输入到卷积神经网络中，对相邻K个词的词向量进行卷积操作获得特征向量表示，当k取多个不同的值，进行多个卷积操作时，可以得到多个不同层次的特征向量表示，对每个特征向量进行Average-pooling、Max-pooling或者Min-pooling采样，取其平均值、最大值或者最小值，最终生成问句和候选答案的特征向量，即分布式语义表示，最后，将问句和候选答案的特征向量进行全连接，将其输入到Softmax层，通过softmax回归函数，将问句和候选答案之间的语义相关性得分转化为概率值，通过概率值的大小对候选答案进行排序，概率值最大即为最佳答案。

4.2 基于BiLSTM的句子对建模

循环神经网络在处理时间序列和语言文本序列时有其独特的优势，而其变本BiLSTM能很好地防止RNN可能存在的梯度消失和梯度爆炸问题，并且可以弥补LSTM只能在一个方向上对序列进行处理的缺陷，因而BiLSTM有着相对广泛的应用，基于BiLSTM的句子对建模也是建立在对问句和候选答案分别进行建模的基础上，如下图4-2所示，是答案选择排序中基于BiLSTM的句子对建模模型：



图4-2 答案选择排序中基于BiLSTM的句子对建模模型

4.3 融合注意力机制的句子对建模

在之前的章节中有介绍到，可以通过BILSTM获取问句和候选答案的特征向量表示，BILSTM的优势在于可以考虑到词与词顺序上的关系而被广泛应用nlp中，但是经过远距离的计算，距离较远的词所起到的作用降低，而在实际中，不同的词语对于句子的语义特征表示有着不同的重要性，本节介将Attention机制融入模型，利用Attention机制生成问句和候选答案中各个词的权重，并根据权重的大小对词语进行不同程度的增强或减弱，从而更好的表达问句和候选答案的语义特征；

本节先后介绍了两种Attention模型，一种是基于问句和候选答案相似性的注意力模型。。。，一种是基于问句和候选答案独立性的注意力模型。。。；

，Attention机制被引入增强关联较强的词权重，降低关联较小的词的权重。采用问题特征与答案特征进行Attention计算，然后将attention计算得到的权值应用到答案特征用以加强对问题较为重要的词的权值，降低对问题不是很重要词的权值。整个算法的流程图如下图2：

4.3.1 基于问句和答案关联性的注意力模型



图4-3： 基于问句和答案相似性的Attetion模型

上图中的和分别是原始的问句和候选答案的词向量矩阵经过BiLSTM后所转化的矩阵表示，和分别为问句和候选答案中词的个数，为BiLSTM中记忆单元的大小，上图中的灰色部分表示Attention模型的计算过程，将Q和A做如式（）的运算，生成注意力矩阵。



其中是模型训练得到的参数，矩阵G代表了问句和候选答案相对于隐藏向量的软对齐得分；接下来，分别以矩阵G的列和行为单位进行最大池化操作，生成向量和，其中，的具体计算如下：

 

 

从式中可以看出，表示问句中的第个词对于候选答案的重要性得分，表示答案中的第个词对于问句的重要性得分，然后通过Softmax归一化操作，将和转化为概率值和，具体计算如下式所示：





根据以上的计算，将得到的和与Q和A进行点乘，如式（）和（）所示，分别得到问句与候选答案最终的语义特征向量和。





4.3.2 基于问句和答案独立性的注意力模型

在上一节介绍的Attention模型中，通过注意力矩阵G的计算以及矩阵在行列方向上的最大池化操作，分别可以得到，问句中的每个词关于候选答案的关联性得分以及候选答案中的每个词关于问题的关联性得分，根据关联性得分以不同的程度加强对彼此较重要的词的权重，这种利用问句和候选答案中词与词之间的关联程度来决定词的重要程度的思路是很方便很直观的，然而，在实际的答案选择排序任务中存在着如下的问题：

1. 在处理众多的候选答案过程中，正确的候选答案只占相对少量的一部分，更多的是要处理不正确的候选答案，对于不正确的候选答案，答案句中最重要的词语片段往往和问句不相同，在这种情况下，关联性和实际的重要性并没有关系
2. 由于本文研究的答案选择排序任务是针对社区问答系统的，所处理的大部分问题都是回答询问建议或观点表达的非事实类问题，而这类问题的正确答案往往和问句之间的相似性较低，所以针对非事实类问题，相似性并不能决定重要性；

综上，需要更多不同的方法决定词语的重要性，以获得更好的问句和候

选答案的特征向量表示。因此，本节在假设问题和候选答案相互独立的基础上提出了一种改进的计算问句和候选答案词语重要性的方法。

本节介绍的方法首先将问句或答案句的词向量矩阵输入BiLSTM网络，得到其特征矩阵，然后再通过一个单独的BiLSTM组合结构来计算输入句子中每个词的重要性大小，并根据重要性加权计算得到句子的最终特征向量表示，这种方法类似于自然语言处理中的self-Attention机制（），不过在本文提出的方法中，句子中词语的重要性计算是基于一个单独的BILSTM组合结构的。（翻译）

4.3.1 基于问题和答案独立性的注意力机制

在针对非事实类答案选择的任务中，具有显著效果的特征训练模型大都是基于BILSTM的，本文提出的方法也是对BILSTM模型的一个变本形式，在假定问句和答案句的独立性的前提下，计算句子片段的重要性，如下图所示，模块1是句子经过BILSTM后的向量表示，模块2是一个基于BILSTM的网络组成计算句子片段重要性权值的过程，模块3是根据句子片段重要性权值进行加权计算得到最终特征表示的过程（原创）



图5：基于问句和答案独立性的Attention模型

在第一个模块中，输入是一个词向量矩阵，代表句子的长度，即句子中词语的个数，代表词向量的维度，经过网络后，分别得到在前向和后向两个方向上的输出矩阵和，在前面的章节介绍过，是的细胞大小，定义是和在行上的连结，并且不经过操作，所以，如（1）式所示：



在第二部分中，以P作为另一个单独的BILSTM组合结构的输入矩阵，得到输出矩阵，的第列是对的第列信息重要性程度的编码，将每一列进行归约运算，转化为一个数值，如（2）所示，然后利用函数将向量按照一定比例大小进行缩放得到，如（3）所示，中的每一个元素代表着对应的第个词语的重要性权值，并且所有的相加等于1。





其中，是用来进行归约运算的网络参数，是句子中所有词语的重要性权值，这里的函数非常重要，它保证了在输入句子的长度更长的情况下，不会累积重要性权值，从而保证了后期运算的可行性和简化性。

第3部分是将句子中所有词语的重要性权值向量和第一部分中未进行池化操作的句子矩阵表示进行加权计算，得到最终的句子特征向量表示，如下图所示：



从上式可以看出，输入句子中不同的片段以不同的强度对句子最终的向量产生影响，即句子中不同的词语以不同的重要性权值进行加权生成最终的句子特征向量表示。

4.3.1实验数据

根据第二章介绍的数据来源及数据抓取操作，得到142063个问题及1285759个答案，本章针对研究内容，从所爬取的数据中剔除没有正确答案的问题以及候选答案不足3个的问答对，然后随机选取28225个问题及其175800个答案作为答案选择排序任务的数据集，然后对这175800个答案进行标记，由于数据量比较大，为了方便高效地完成标记工作，将爬取网站数据时带有最佳答案标签的答案或点赞数最多的答案作为问题对应的最佳答案，并以1作为标记，其余的答案以0作为标记；并且将这个数据集以3:1:1的比例分为训练集（Train）、验证集（Dev）、测试集（Test）三部分，具体数据集统计信息如下图所示：

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Train | Dev | Test |
| Number\_Q | 16935 | 5600 | 5690 |
| Number\_A | 104997 | 30400 | 40403 |
| Length\_Q |  |  |  |
| Length\_A |  |  |  |

表4-7 数据集统计信息

实验环境

Tensorflow

Python

CBOW-pos&clu训练得到的200维的词向量

评价指标

本论文主要是针对候选答案选择排序的研究，实验的目的在于将问题对应的所有候选答案依照和问题的语义相关程度进行排序，选出相关性得分最高的答案，即为问题对应的最佳答案。因此，本文的评价指标采用平均准确率均值(MAP,Mean Average Precision)和平均倒排名(MRR,Mean Reciprocal Rank),这两种评价指标都是信息检索中比较常用的评价方法，MAP表示的是每条查询内容的平均准确率的均值，反映了系统在与查询内容相关的所有文档集合上的性能指标，MRR表示的是一个查询集合的倒排名平均值，反映的是系统返回的正确答案所处的位置，在实际的系统评价中，MAP和MRR这两个指标的值越高，就表示检索结果的平均准确率越高，问题对应的最佳答案的位置排序越靠前，系统的性能也就越好，MAP和MRR的具体计算公式如下：







其中：

在信息检索中表示查询内容的个数，在本研究中代表的是所有问题的个数，表示在保证每个问题的所有正确都包含在集合内的前提下，问题集合中第个问题对应的检索排序结果的平均准确率，表示第个问题所对应的候选答案的个数，表示在个候选答案中，第个候选答案经过排序后在候选答案序列中的位置，表示的是与第个问题最相关的答案在候选答案序列中的位置；

实验设置

候选答案选择排序可以被视为一个二分类问题，假设问题集合中的每一个问题都对应一个候选答案序列及其标签序列，

如果第个候选答案是问题的正确答案，则否则，形式化的表示为，给定一个问题的向量表示和其第个候选答案的向量表示，那么这个候选答案为正确答案的概率即的标签的概率为：



式中的偏置系数和转换矩阵都是模型参数，这个公式也可以理解为，给定一个候选答案句，通过转换操作“生成”一个问题，然后通过点积操作度量生成问题和给定问题之间的距离，sigmoid函数将问题和答案的语义相关性得分转换为0到1之间的概率值，并依此概率值的大小进行排序。

神经网络训练的过程，是以损失函数(loss function)作为方向性指导，对相关参数进行调整的过程，对于损失函数的收敛特性，期望的是当误差越大的时候，收敛的速度即学习的速度应该越快。

在本文的研究中，模型的训练使用交叉熵损失函数，其具有非负性的特性，并且当真实输出值与期望输出值接近时，其函数值趋近于0，模型训练的目标就是最小化交叉熵损失函数，如下式所示：



在损失函数中加入了惩罚项以防止过拟合，是惩罚项系数，是包含和的参数集，根据链式法则，将损失函数分别对参数、求偏导、，并在反向传播的过程中，通过随机梯度下降法来更新模型的参数，使模型最终趋于收敛。



实验设置

在本章的前三节中，介绍了基于多种深度神经网络的答案选择排序模型，本节将分别对每种神经网络结构的超参数设置情况做以说明。

词向量作为第三章重点介绍的内容，它不仅是输入，

实验结果分析

4.4 实验设置

4.3.1 数据集介绍及预处理 1

4.3.2 模型训练初始化及优化 2

4.4实验结果分析与对比 3

4.5 本章小结 0.5