1. **基于深度学习的答案选择排序模型**

本章在第三章词向量计算的基础上，构建了基于深度学习的答案选择模型，基于神经网络分别对问句和候选答案进行句子对建模，得到句子语义特征向量，然后通过最小化交叉熵损失函数对模型进行训练，最终得到问题和每个候选答案的语义相关性得分，达到对候选答案进行选择排序的目的；本章着重研究句子对的建模问题，针对答案选择排序问题，对神经网络的组合结构及模型参数做出对比和改进，并在相关数据集上进行实验，并从信息检索的角度对比了CNN模型、BILSTM模型、融入Attention机制的模型以及改进的Attention模型在答案选择排序任务上的性能；

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，以图示的方式进行表示，得到答案选择排序中基于CNN的句子对建模模型，如图4-1所示：



图4-1：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。

将上述过程进行具体化描述，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：

**** （4-1）

 （4-2）

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。

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，如图4-2所示，是答案选择排序中基于BiLSTM的句子对建模模型;

在第二章已经介绍了循环神经网络及其改进后的长短期记忆网络处理文本序列信息的原理及过程，本章针对问句和候选答案，首先进行tokenize分词操作，然后通过look up table查找每一个词语对应的向量，生成词向量矩阵，将其作为网络的输入，每一个词语对应于BILSTM的每一个时间步time\_step，每个词语的词向量对应每个时序的输入，如2.4节所介绍，词的部分信息通过遗忘门（forget gate）、输入门(input gate)选择性地从细胞状态中被丢弃或选择性地加入到细胞状态中去，最后，由输出门（output gate）输出词的信息，在传统的对于BILSTM输出的处理中，会将经过传递后，序列中最后一个time\_step即最后一个词语的状态输出作为最终的句子特征向量表示，而经过不断地尝试，研究者们发现，将每一个time\_step下每个词语的隐藏状态进行不同的池化（pooling）操作，会有更突出地效果，本文就是使用这种方法，将每一个time\_step下前向和后项两个方向上的隐藏状态进行连结，然后进行max-pooling操作，取每一维的最大值组成输出向量作为整句话的语义特征向量；



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于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，考虑到BILSTM在隐藏层计算单元的限制，Attention机制被引入模型来计算句子中不同词语的权重，并对不同的词语进行不同程度的增强或减弱；

4.3.1 传统Attention模型

如下图4-3所示，在进行Max-pooling或Average-pooling之前，每一个BILSTM输出向量会乘以一个softmax权重，这个权重是由BILSTM获得的问题向量决定的，假设将答案句第t个时间步的输出向量记为,问句编码记为，更新的为，表示加入了Attention计算后得到的第t个时间步下的权值，具体的计算过程如下式所示：

 （4-3）

 （4-4）

 （4-5）

其中，，是注意力参数，注意力机制的应用使得某些词获得了更大的权重，而在上面的例子中，权重的计算是通过问题信息获得的，在答案选择任务中，直接利用基于Attention的向量表示度量问题和答案之间的距离，可以通过问句信息有效地区分正确答案和不正确的答案。



图4-3 Attentin机制应用于问句和候选答案的句子对建模

此种Attention模型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增强候选答案中和问题关联较强的词的权重，降低和问题关联较小的词的权重。

4.3.2 APN模型

文献（43）中提出的关注池化神经网络（APN，Attentive Pooling Network）同样也是Attention机制的在自然语言处理中的一种表现形式，图4-4是APN的结构，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（4-6）的运算，生成注意力矩阵。

 （4-6）



图4-4： 基于APN的答案选择排序模型

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：

  （4-7）

  （4-8）

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：

 （4-9）

（4-10）

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。

 （4-11）

 （4-12）

4.4 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。

如下图所示：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（4-13）式所示：

 （4-

图4-5：基于问句和答案独立性的Attention模型

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。

 （4-14）

 （4-15）

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：

 （4-16）

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，得到百度知道数据集（ZhiDao）142063个问题及1285759个答案，法律领域数据集（law）93882个问题及892379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ZhiDao中28225个问题及其175800个答案作为答案选择排序任务的数据集，选取Law中19972个问题及95564个答案，然后分别对这些答案进行标记，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；并且将这两个数据集以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表所示:

表4-1 数据集统计信息

|  |  |  |
| --- | --- | --- |
| Data | Zhidao Law | |
| Train Dev Test | Train Dev Test |
| num\_Q | 16935 5645 5645 | 11982 3995 3995 |
| num\_ A | 105480 35160 35160 | 57338 19113 19113 |
| Average length\_Q | 36 33 34 | 29 32 31 |
| Average length\_ A | 150 160 150 | 140 155 150 |

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：

 （4-17）

 （4-18）

 （4-19）

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：

 （4-20）

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：

（4-21）

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，它不仅是神经网络的输入，更是神经网络模型的一种重要的超参数，本章的研究内容是以第三章的研究为基础，并以第三章提出的CBOW-pos&clu模型训练出来的词向量作为本章答案选择排序模型的输入，词向量的维数依照第三章的训练结果取200维；并且，根据之前研究中（）的经验所得，将训练方式设置为static，即不需要在模型训练过程中对词向量重新训练。

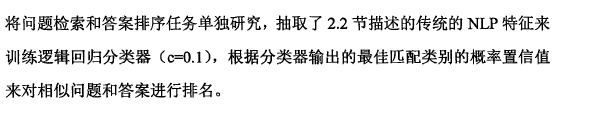
在基于CNN的句子对建模中，考虑到问题和候选答案的句子长度有所差别，某些参数的设置也相应有所差异，问题和候选答案分别各用一组CNN结构，问题的滤波器窗口Filter Window分别设置为2,3,4几个值，候选答案的滤波器窗口设置为3,5,6，并且都以50为间隔设置不同的Feature Map数量，对比Feature Map的数量从100到600时模型性能的好坏；池化层设置为最大池化(max-pooling);

针对基于BiLSTM的句子对建模，问题和候选答案共用同一组网络，共享同一组参数，Memory Size的大小以50为间隔，设置为从100到600的值，探究不同的Memory Size对模型性能的影响；

在融合Attention机制的句子对建模中，本文是在BiLSTM模型的基础上加入Attention机制，并以BiLSTM模型单独进行句子对建模时性能最好的参数作为融合Attention机制的模型的标准超参数；

为了防止过拟合，在全连接层加入了Dropout，Dropout的参数为0.5，表示以0.5的概率随机地舍弃一部分连接，在sigmoid分类时加入L2正则项，正则化权重设为0.0001；训练的时候，SGD的mini-batchsize是60，这样，模型能够在100个epoches内收敛，同时，神经网络的学习速率设置为0.005。

基线系统的设置



4.6 实验结果分析





图4-6 基于CNN的答案选择排序模型在测试集上的性能表现

图4-6显示了CNN模型设置不同的Feature Map时，在数据集上的不同表现结果，上侧图为ZhiDao数据集上的性能表现，下侧图为Law数据集上的性能表现，从这两幅图中可以看出，在数据集规模相差不是很大的情况下，Law数据集上的性能表现整体好于ZhiDao数据集，

同时，还可以看出，不管是哪个数据集，首先随着Feature Map数量的增加，模型的性能呈现为不稳定上升的趋势，在Feature Map为200左右时，模型的性能达到最好，之后随着Feature Map的继续增加，模型的性能开始不稳定下降，之后再没有出现最高点；





图4-7 基于BILSTM的答案选择排序模型在测试集上的性能表现

图4-7显示了BILSTM模型设置不同的Memory Size时，在数据集上的不同表现结果，上侧图为ZhiDao数据集上的性能表现，下侧图为Law数据集上的性能表现；

同样的，从上面两幅图中可以看出，在数据集规模相差不是很大的情况下，Law数据集上的性能表现整体好于ZhiDao数据集；出现这种结果的原因可能是特定域的数据集内容相对集中，句子的主题相对明确，语义特征的捕捉相对准确，也可能是Law数据集在第三章词向量计算部分就已取得相对较好的性能，进而对本章答案选择排序模型的性能产生了积极的作用；

另外一点是，类似于CNN模型，BILSTM模型中随着Memory Size的增加，模型的性能呈现为先上升后下降的情况；这种变化趋势表明，在基于神经网络的句子对建模中，句子特征向量表示的维度并不是越大越好，当维度超过一定限制时，会因过大的维度将特征信息稀疏化从而影响模型的性能；

对比上面的两组图，能看出基于BILSTM的模型性能优于CNN模型，虽然在句子特征向量表示的维度过大时两者的性能趋于一致甚至略差于CNN模型，但当Memory Size取200或300时，BILSTM模型表现出了最优的性能，相比CNN模型略胜一筹；这就再次证明了循环神经网络在处理序列信息时独特的优势，相比于CNN通过卷积捕捉连续几个上下文词语的信息，BILSTM是通过信息的传递与取舍捕捉整个句子序列的信息，并且从前后两个方向同时进行，捕捉每一个点完整的过去和未来的上下文信息；从以上的分析可知，在答案选择排序任务上，基于BILSTM的句子对建模要优于基于CNN的句子对建模；

Attention模型的可视化分析

本节将不同Attention模型在答案选择排序过程中对问句和候选答案进行的不同Attention 计算的结果进行可视化分析，基于不同的Attention模型，答案选择排序的性能也就不同，同样的问题对应的最佳答案的预测结果也会呈现出不同；在本节的研究分析中，随机选取测试集中的三个问题、、及其基于不同的Attention模型所预测的最佳答案，问题如下所示：

：公司已注销，原注册商标是否还属于原公司所有？怎么办理

转让手续？（from Law）

：版权法和著作权法一样吗？有区别吗？(from Law)

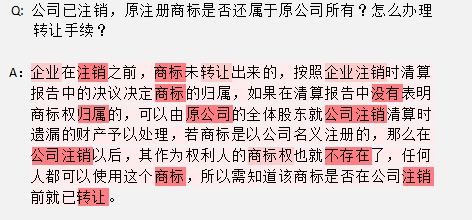
：蓝牙遥控器使用时如何操作？(from ZhiDao)

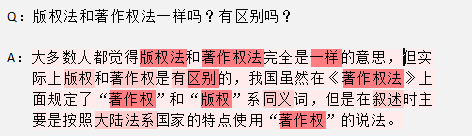
这三个问题分别在基于不同的Attention模型下预测的最佳答案结果如表4-3所示。

表4-3 不同的Attention 模型下最佳答案的预测结果

|  |  |  |  |
| --- | --- | --- | --- |
| Attention Model | Q1 | Q2 | Q3 |
| 传统Attention | 预测正确 | 预测不正确 | 预测不正确 |
| APN | 预测正确 | 预测不正确 | 预测不正确 |
| self Attention | 预测正确 | 预测正确 | 预测正确 |

图4-8、图4-9、图4-10分别针对不同的问题对比分析了了不同的Attention模型进行Attention计算的可视化热图；显示了问题答案对通过Attention模型后计算得到的词语片段权重及分布结果，其中颜色越深，表明计算所得该词的权重越大，该词对于句子的语义特征表示越重要，从这些可视化分布中也能看出不同的Attention模型对于句子语义建模的不同影响和作用；





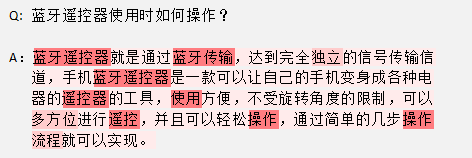
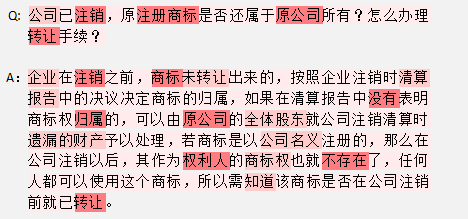
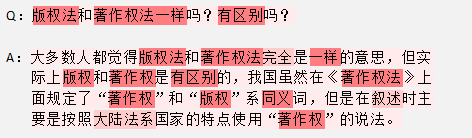


图4-8 传统Attention模型在预测集上的可视化热图





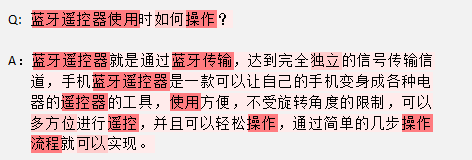
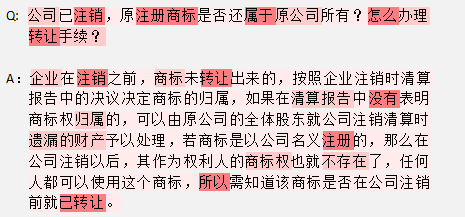
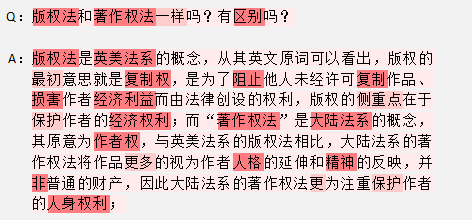


图4-9 APN模型在预测集上的可视化热图





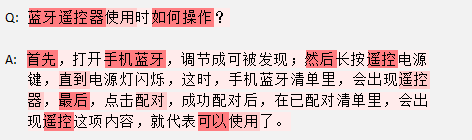


图4-11 self Attention模型在预测集上的可视化热图

（c）self Attention--预测正确

图4-8 传统Attention、APN、self Attention在上的可视化

从上图中可以看出，APN模型在处理问答对时，问题Q1中出现的关键词“公司”、“注销”、“注册”、“商标”、“转让”等被赋予很高的权重，该模型认为，在A1中那些和Q1关键词具有很高相似性的词会被认为是对A1的语义特征表示更重要的词，会被赋予更高的权重，从图中可以看出，A1中包含多个与Q1中关键词相似的词语，这些词被赋予了很高的权重；

上图中的两组问题候选答案对均取自实验数据集中的测试集部分，这两组问题候选答案对中的候选答案都是被标记为“1”的该问题的最佳答案，APN模型对第一组问答对的预测取得了正确的结果，但对第二组问答对没有预测正确，而self-Attetion模型在两组问答对上均获得了正确的预测结果；

上图分别显示了两组问题候选答案对Q1-A1、Q2-A2通过两种Attention模型后词语片段权重大小的计算及分布结果，其中颜色越深，表明计算所得该词的权重越大，该词对于句子的语义特征表示越重要；从上图的可视化分布中也能看出两种Attention模型对于句子语义建模的不同影响和作用；

对于第一种Attention模型APN，在处理第一组问答对Q1-A1时，问题中出现的关键词“公司”、“注销”、“注册”、“商标”、“转让”等被赋予很高的权重，该模型认为，在A1中那些和Q1关键词具有很高相似性的词会被认为是对A1的语义特征表示更重要的词，会被赋予更高的权重，从图中可以看出，A1中包含多个与Q1中关键词相似的词语，这些词被赋予了很高的权重；而在第二组问答对Q2-A2中，Q2中出现了关键词“版权法”、“著作权法”、“区别”，这几个词分别被赋予较高的权重，对于A2，被赋予较高权重的词语仅仅是与Q2中关键词有较高相似性的“版权法”、“复制权”、“著作权法”、“作者权”等，并没有对能够有效表征A2语义特征的词，如“阻止”、“复制”、“经济利益”、“保护”“经济权利”“人格”、“精神”、“人身权利”等词赋予较高权重，显然，APN模型在处理类似Q2-A2这样的问答对时，只是片面的认为，相似的词语应该有更高的权重，而这样并不能全面有效地得到句子的语义特征表示；

然而，第二种Attention模型self-Attention对两组问答对的处理都是建立在问题和候选答案相互独立的基础上，如图所示，在Q1中，关键词“公司”、“注销”等被赋予了较高的权重，A1中存在同样的关键词，也被赋予了较高的权重；在Q2中，关键词“版权法”、“著作权法”等被赋予较高的权重，而在A2中，关键词有所不同，一些和“版权法”、“著作权法”并没有较高相似性的词“经济权利”、“人格”、“精神”、“人身权利”等也被赋予了较高的权重，模型认为，这些词的重要性并不是建立在其和Q2有没有相似性的基础上，而是建立在其本身对于A2的语义表示所起的作用上，实际中也能看出，这些词也确实更能体现A2的语义特征；这也说明，self-Attetion更能关注到问句和候选答案中的重要信息，从而赋予其更高的权重；

上表列出了列出了基于CNN、BILSTM以及融合了Attention机制的句子对建模在测试集上的结果，其中，CNN和BILSTM的结果值是在Feature Map和Memory Size分别取300和200，模型性能达到最好时的最佳结果值，根据上表的数据，可以看出，

|  |  |  |
| --- | --- | --- |
| Model | Dataset | |
| ZhiDao Law | |
| MAP MRR | MAP MRR |
| Ranking SVM |  |  |
| CNN | 78.33 86.52 | **78.57 86.63** |
| BILSTM | 78.98 87.21 | 79.45 87.39 |
| BILSTM+APN | **78.30 86.29** | 81.06 87.49 |
| BILSTM+  self Attention | **80.25 87.56** | **82.18 88.36** |

实验总结

本章小结

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；