面向多片段答案的抽取式阅读理解模型

苏立新[1),2),3)] 郭嘉丰[1),2),3)] 范意兴[1),2)] 兰艳艳[1),2),3)] 徐 君[1),2)] 程学旗[1),2),3)]

[1)](中国科学院网络数据科学与技术重点实验室 北京 100190)

[2)](中国科学院计算技术研究所 北京 100190)

[3)](中国科学院大学 北京 100049)

摘 要 随着搜索技术的发展,抽取式阅读理解已经成为搜索引擎中重要的组成部分.给定问题和文本,抽取式阅读理解任务要求从文本中定位出问题的答案.已有工作仅考虑答案片段由文本中的一个片段组成的情况,因此把该问题建模为输入问题和文本,预测出两个文本中的位置索引去指示答案的起始和结束位置.然而现实应用中存在大量问题其答案往往由文本中一个或多个片段组成,想要回答该问题需要从文本中定位出若干的文本片段,而不再是单一片段.已有的阅读理解模型研究主要关注在模型底层结构的设计,对于多片段答案的情况未予考虑,导致已有模型无法从文本中抽取多个答案片段去回答问题.本文提出面向多片段答案的抽取式阅读理解模型 BERT-Boundary,该模型采用预训练的 BERT 作为底层结构进行文本和问题的理解.BERT 通过自我注意力机制和前向神经网络对文本和问题进行编码表示,同时利用在大规模无监督语料上进行 BERT 模型参数的预训练达到更强的文本理解.利用新颖的边界序列标注方式去建模一段文本中多个答案片段,模型对答案的起始位置和结束位置分别进行序列标注,对每个词进行二分类,判断其是否是答案的起始位置或者结束位置,并通过简单有效的序列标注方式进行答案片段的解码.BERT-Boundary 结合了 BERT 的文本理解能力和边界序列标注的多片段建模能力.我们在构造的大规模多片段答案的阅读理解数据集上进行详尽地实验和分析,实验结果表明,BERT-Boundary 的性能比基线方法取得一致的提升.我们进一步在不同答案片段长度和答案片段数量上比较我们的模型和基线方法,实验数据表明,我们的方法比基线方法取得一致的提升.我们的代码公开发布在 https://github.com/lixinsu/multi_span.

关键词 阅读理解;多片段答案;问题系统

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.00856

A Reading Comprehension Model for Multiple-Span Answers

SU Li-Xin[1),2),3)] GUO Jia-Feng[1),2),3)] FAN Yi-Xing[1),2)] LAN Yan-Yan[1),2),3)] XU Jun[1),2)] CHENG Xue-Qi[1),2),3

[1)](*Key Laboratory of Network Data Science and Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*University of Chinese Academy of Sciences, Beijing* 100049)

Abstract With the development of the Internet, extractive question answering became an important component in the modern search engine systems. Given the question and the related passage, extractive reading comprehension models aim to find answer texts from the related passage to answer the question. Existing works only considered the setting where the answer to the question is comprised of a single text span from the passage, and they regarded the problem as taking the passage and the question as input and predicting two position indices in the passage to indicating the answer. However, in fact there are many questions whose answers are comprised of many text spans from the related passage, and the models needs to locate multiple spans from the

收稿日期:2018-10-17;在线出版日期:2019-08-19.本课题得到国家自然科学基金重点项目(61425016,61472401,61722211,20180290)、中国科学院青年创新促进会(20144310,2016102)、国家重点研发计划(2016QY02D0405)资助.**苏立新**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为信息检索、自动问答.E-mail:sulixin17b@ict.ac.cn.**郭嘉丰**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为信息检索与数据挖掘.**范意兴**,博士,助理研究员,中国计算机学会(CCF)学生会员,主要研究方向为信息检索与文本挖掘.**兰艳艳**,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为统计机器学习、排序学习和信息检索.**徐 君**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为信息检索与数据挖掘.**程学旗**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为网络科学、互联网搜索与挖掘和信息安全等.

passage to answer the question, rather than the single answer span. Existing works for reading comprehension models mainly focused on the design of the lower layer for text understanding and lacked of exploration on the multi-span answers. In this paper, we proposed BERT-Boundary model which uses the pre-trained BERT as lower layer. Specifically, BERT employs the self-attention mechanism and the fully-connected feed forward network to encode the related passage and the question, and it used the pre-trained parameters which are trained on the large-scale unsupervised corpus in advance. Our proposed model employs an answer-boundary labeling model for locating multiple answer spans. Specifically, it conducts sequence labeling for the start positions and the end positions of the answer spans over the related passage and a classification is conducted over each token of the passage. BERT-boundary combines the power of text understanding and multi-span locating. We conduct extensive experiments on the newly constructed multi-span answer dataset. The experimental results demonstrate the effectiveness of our model and a clear improvement on $F1$ measurement over existing models. First, we compare our language understanding layer of our model with RNN-based lower text understanding layer, and we find that our BERT-based model achieve better result. Second, we also compare the answer component of our model with content-based method which models whether a token is inside an answer span or not, and the result demonstrates that our boundary labeling strategy can locate answer spans more accurately than the content-based method. Finally, we compare our model with original BERT model based on the SQuAD data set, and we can find that our model can locate multiple answer span from the passage, meanwhile our model can still process single-span answer in usual reading comprehension data set. We further analyze the performance with the number of the answer span and the length of the answer span. As the increase of the number of the answer span, our model performance drop slowly while performance of previous BERT model drops quickly. As the increase of the length of the answer span, we can find that our model performs well on locating longer answer. However other baseline models cannot handle the long answer texts, especially for the content-based method, its performance drops as the length of the answer increases. We conclude that our model achieves consistent performance gain compared to existing baseline methods. Our code is publicly available at https://github.com/lixinsu/multi_span.

Keywords　reading comprehension; multiple-span answers; problem system

1　引　言

随着互联网的发展，数据规模快速增长，帮助用户在大量的数据中快速精准地找到有用的信息是一个长期有挑战的任务. 机器阅读理解可以在搜索结果页基础上帮助用户阅读文本找到精确的答案. 抽取式阅读理解的形式为给定一段文本和问题，模型从文本中找出片段去回答问题. 搜索引擎中大量信息类查询可以通过阅读理解技术帮助用户快速获得答案而不再是一个排序的文档列表，这样可以极大提升用户体验. 尤其在移动场景下，屏幕尺寸受限，精确简洁的答案更能提升用户的体验.

然而现有的阅读理解模型和相关数据集存在一定限制，限定答案仅由文本中单个片段组成. 这限制了阅读理解模型的应用. 在实际的应用中，存在大量问题其答案不能由文本中的一个片段回答，而是由文本中多处片段来共同回答. 一个例子如图 1 所示，问题是“和中国接壤的国家有哪些?”，这个问题的答案包含多个实体，且分布在文本中. 回答该问题需要从文本中抽取出多个片段. 以往研究中答案仅由一个片段组成，如图 1 中问题“湖南的省会是哪里?”. 本文将问题的答案可能是文本中一个或多个片段的形式称为多片段答案的阅读理解. 该任务定义作为单片段答案的阅读理解的超集，能覆盖更多实际问题，更加符合实际的应用需求.

已有的工作主要针对答案由单片段组成的情况，如 SQuAD 数据集[1]、TriviaQA 数据集[2]. 请注意多文本阅读理解（如 TriviaQA）和本文研究的多片

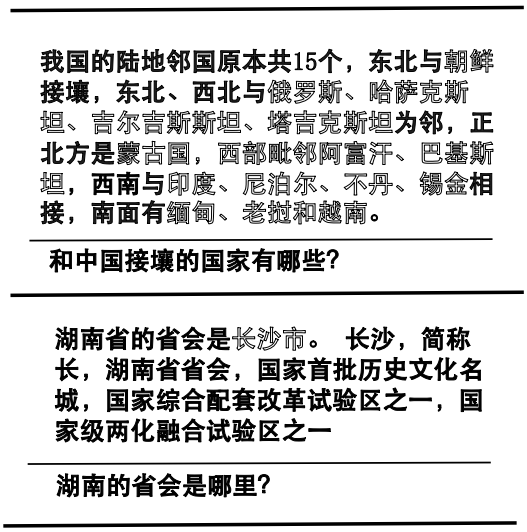

图 1 答案有一个片段组成和答案由多个片段组成

段答案的阅读理解的区别，前者强调用于寻找答案的文本有多个，后者强调答案由文本中一个或多个片段组成. 已有的模型仅考虑了如何从文本中抽取一个片段，并将问题转化为预测答案片段的起始、结束位置. 如图 1 所示的例子“湖南的省会是哪里”，模型仅需要输出两个位置，长沙市对应的起始位置和结束位置. 然而在多片段答案的情形下，答案片段数量不再是一个，且不同问题对应的答案片段数量也在变化. 已有的数据集不能支持多片段答案的研究，同时已有的方法也不能完成多个片段的位置建模.

本文提出一种新的多片段答案阅读理解模型 BERT-Boundary，该模型的底层结构基于预训练的 BERT[3] 结构. 在答案输出层通过答案边界序列标注的方式建模多个答案片段的位置. 具体来说，在答案的输出层本文采用对文本每个词项进行分类，分类的类别是该词项是否是答案的起始位置/结束位置. 同时在模型预测过程中，本文采用新的解码算法，将起始位置和结束位置进行配对，解码出多个答案片段. 同时本文的方法可以通过调节分类的阈值，平衡返回的答案的精确率和召回率以适应实际应用场景中的不同需求.

本文在新收集的大规模多片段答案的数据集合上进行实验，验证所提模型的有效性. 同时在 SQuAD 数据集合上和原来 BERT 模型的性能对比，说明我们的方法在建模多个答案位置的同时不大幅降低单片段答案上的性能. 另外我们比较基于 RNN 底层结构和基于 BERT 结构的方法，证明选用 BERT 作为底层结构对于整体性能的优势. 同时我们对比基于 BERT-Content 的方法，该方法是解决多片段答案的一种朴素方法，通过实验验证了所提 BERT-Boundary 是解决多片段答案的较优方法.

本文的主要贡献如下：

(1) 提出一种新的多片段答案的阅读理解模型用来解决答案若干片段组成的情形. 模型结合了 BERT 和边界序列标注，有效定位多个答案片段.

(2) 本文在两种底层结构 RNN 和 Transformer 以及两种多片段答案抽取方案-Content 和-Boundary 上分别进行实验. 验证了 BERT-Boundary 是较优的设计.

(3) 本文进行了详尽的实验和分析，表明所提模型的有效性，在数据集上取得一致的提升.

2 相关工作

近年来机器阅读理解技术快速发展，机器阅读理解在计算资源和深度学习的推动下展现出显著效果. 本节从任务形式和模型结构两个维度对已有工作进行回顾.

2.1 任务形式

阅读理解任务从问题形式上可以分为三种主要类型：填空式阅读理解，选择式阅读理解，抽取式阅读理解.

填空式阅读理解的任务形式是给定一段文本和一个问题，通过阅读文本将问题中的缺失部分补充完整，代表性的数据集有 CNN/Daily[4]、CBT[5] 等. 数据的构造过程是将新闻的摘要部分作为问题，将其中的部分实体移除，要求根据文章内容预测出缺失的实体内容. 为了防止填空问题退化成语言模型问题，相关实体一般被替换成实体标识符进行匿名化. 填空式阅读理解的特点是训练数据构建成本低，可以大量自动化构建.

选择式阅读理解任务定义为给定文本和问题及若干候选答案选项，要求机器通过阅读文本和问题，从候选答案选项中选择出正确的答案选项. 典型的数据集合有 MCTest[6] 和 RACE[7]. 其中 RACE 数据集构造过程采用的是现有中学生和高中生英语阅读理解题目，这些题目是由专家命题用来考察学生们的阅读理解能力，所以其中包含多种多样的推理和语言学现象. 选择题类阅读理解可以很好的考察机器的文本理解和推理能力，近年来新发布的很多数据集都采用此形式，如 CommonsenseQA[8]、RecipeQA[9] 等，它们进一步考察机器的常识推理能力和多模态联合推理能力.

抽取式阅读理解任务定义为给定一段文本和问题，要去从文本中抽取内容回答该问题. 该类型阅读

理解的典型数据集合如 SQuAD[1]、TriviaQA[2]、SearchQA[10]等. 其中 SQuAD 通过人工阅读一段维基百科文本,构造若干问题-答案对,答案作为一个连续片段存在于维基百科文本. TriviaQA 和 SearchQA 的构造过程是首先收集问题和答案对,其中问题多为事实类问题,然后检索获得相关文本,保证答案出现在检索得到的相关文本中,检索结果未出现答案的样本被过滤掉. SQuAD、TriviaQA 和 SearchQA 保证答案作为单个连续片段出现在文本中. DuReader[11]和 MS MARCO[12]两个数据集通过搜索引擎的用户查询作为问题,搜索引擎的结果作为文本,利用众包用户去回答问题,其产生的答案大部分存在于文本中,部分不能直接找到答案. CoQA[13]和 QuAC[14]两个数据集把抽取式阅读理解引入到对话场景中,其数据集的标注过程为每条数据雇用两个众包工人,一个人连续提问题,另一个人根据文本进行回答,大部分答案(除去是否类问题)可以通过从文本中抽取一个片段进行回答.

我们可以看到抽取式机器阅读理解模型更加侧重问答的应用,在搜索和对话场景都有探索. 其可以作为搜索引擎的延伸,帮助用户阅读结果,定位精确答案以节约用户时间. 而填空和选择类更加侧重于考察机器对于文本的理解. 本文从实际应用出发,扩展原有抽取式阅读,我们考虑答案片段不再仅仅由文本中一个片段组成的情况. 而是根据实际的问题和文本,抽取若干片段回答问题. 这样的设定使得抽取式阅读理解可以解决更多的问题,产生更大的应用价值.

2.2　模型结构

本节我们综述阅读理解的模型结构. 不失一般性,阅读理解模型从逻辑上可以划分为阅读模块和答案预测模块两个部分. 阅读模块主要是对输入的问题和文本进行理解表示,答案预测模块是根据阅读模块生成的表示进行答案的预测.

阅读模块的核心在于对齐文本和问题中的信息,在问题信息指导下进行文本的表示. 已有工作如 Wang 等人[15]提出通过带有注意力机制的 LSTM 结构进行文本和问题信息的对齐,然后将对齐后向量送入 LSTM 编码得到最终的文本表示. Seo 等人[16]提出双向注意力机制,使得文本通过两次注意力机制对齐到问题,更加明确文本中的哪些词是和问题相关的. RNET[17]中提出在文本上应用自我注意力机制,通过该机制文本中的每一个词可以获得全文的内容,提升对长文本的理解,并且引入了门机制过滤信息. FusionNet[18]和 SLQA[19]等模型对文本和问题首先进行多层级的表示,然后将文本和问题在不同层级应用注意力机制进行对齐,并设计细粒度的混合函数去混合注意力向量和层级表示. 迁移学习研究中的上下文感知的预训练词向量也在文本理解上显示出卓越的性能,如 Cove[20]和 ELMo[21]在用在各种阅读理解模型 RMReader[22]、SAN[23]上替换原有的词向量都获得了显著的性能提升. BERT 模型引入多层预训练的 Transformer,借助 Transformer 的表达能力和大规模语料在 SQuAD 上取得最好性能. 本文所提的模型采用 BERT 作为阅读理解模型的底层结构,同时也使用一种基于 RNN 的底层结构替换 BERT 进行对比实验.

针对答案预测模块,不同的任务类型需要不同的答案预测结构. 填空类型阅读理解需要模型预测缺失的内容是词表中的哪一个词. 具体来说模型根据阅读模块产生的表示向量映射到词表空间分布[24]. 针对选题式阅读理解,答案预测模块需要预测多个候选答案的分数分布. 选择概率最大的选项作为最终答案. 具体的模型如 Wang 等人[25]用文本和问题与每个候选选项进行匹配打分.

针对问答式阅读理解,已有工作中假定文本中只有一个答案片段,那么就可以转化为对一对起始结束位置的建模. 模型需要根据阅读模块输出的文本表示预测两个概率分布,一个代表起始位置的概率分布,一个代表结束的概率分布. 上文提到的大多数模型如 Wang[15]、Seo[16]、FusionNet[18]和 SLQA[19]等都是采用在阅读模块输出的表示上应用指针网络或者分类网络,产生起始位置和结束位置的概率分布. Li 等人[26]采用序列标注的方式去抽取答案,标注文本中的每一个词是否是答案中的词. R^3[27]、Lin 等人[28]、VNET[29]、CascadeQA[30]等针对 SearchQA、MSMARCO、DuReader 从多个文本抽取答案的情况,模型中引入排序机制,对多个文本进行排序,或者是对文本中抽取出的多个答案进行排序. 这些模型组合排序模型和单片段阅读理解模型去解决多文本阅读理解.

本文所提多片段答案的阅读理解是抽取式阅读理解的延伸. 我们需要根据需要从文本中定位出多个片段组成答案. 上文提到的模型无法满足这些要求,它们都是从一个文本中抽取单个片段. 注意 Li 等人[26]采用的序列标注方式可以改造为多片段答案预测的模块,我们改进其解码算法,将其作为一种对比基线. 本文将采用该答案预测结构的模型标注为-Content,由于其直接对答案内容进行标注.

3 模型方法

本节将介绍多片段答案的阅读理解的形式化定义.然后对所提的多片段答案阅读理解模型 BERT-Boundary 的结构进行描述,我们将其分为阅读模块 BERT 和答案预测模块 Boundary 两部分去介绍,同时我们也介绍了其他可选择的方案如基于 RNN 的阅读模块和基于 Content 答案预测模块,用于和我们方法做对比.最后我们对预测阶段的顺序匹配解码算法进行介绍.

3.1 问题定义

多片段答案的阅读理解可以定义为:给定一个问题和一段相关文本,从文本中找出若干连续片段回答该问题.$q_{1:n}$表示问题中的词序列 $q_1,\cdots,q_n$,$p_{1:m}$表示文本词序列 $p_1,\cdots,p_m$,其中 q_i,p_j 均为词对应的词典中的索引.给定 $q_{1:n}$和 $p_{1:m}$,模型需要从文本词序列中找出 k 个片段回答该问题 $p_{s_1:e_1},p_{s_2:e_2},\cdots,p_{s_k:e_k}$ 回答该问题,其中 s_i,e_i 为第 i 个答案片段的边界,k 为正确答案的片段数量,且 $k\geq 1$.注意 k 需要根据问题 q 和文本 p 确定.

3.2 基于 BERT 的阅读模块

本文基于预训练的 BERT 作为模型底层结构去交互输入的文本和问题,生成文本的表示,该表示用于答案预测模块定位答案.

首先介绍 BERT 是如何将离散的词符号映射到向量.预处理时 BERT 对文本和问题进行分词后得到若干词项,注意词项不是常规的单词,BERT 分词后输出的是子词(sub-word),这些词项作为模型的输入.问题和文本拼接成一个序列输入到模型中.每个词项映射到一个词的内容向量 $\boldsymbol{CE}(w_i)$,每个词项对应一个可学习向量,一个位置向量 $\boldsymbol{PE}(w_i)$,词所在的绝对位置对应一个向量和一个段向量 $\boldsymbol{SE}(w_i)$,表示词来自文本还是问题.把以上三者相加得到最终词向量表示

$$\boldsymbol{E}(w_i)=\boldsymbol{CE}(w_i)+\boldsymbol{PE}(w_i)+\boldsymbol{SE}(w_i).$$

BERT 中用于编码和交互的结构采用堆叠式的 12 层相同的 Transformer 结构,如图 2 所示.每一层结构中包含自我注意力机制和共享的前向神经网络.我们对这个两个子结构分别应用残差连接和层正则化,每个层的输出如下:

$$LayerNorm(x+Sublayer(x)),$$

其中子层($Sublayer$)可以是多头注意力机制和前向神经网络.下面我们展开介绍这两种子层.

多头注意力机制不同于以往的注意力机制,其

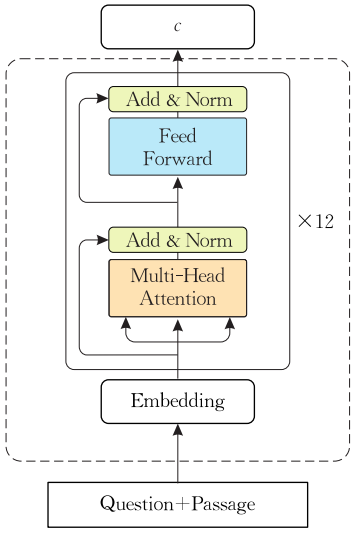

图 2 BERT 的底层结构

在多个子空间中计算两个向量的相似度.多头注意力机制首先将数据投影到 h 个空间中,h 表示注意机制头的数量.我们利用 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 表示注意力机制中的输入,那么映射到一个空间中的向量为 $\boldsymbol{QW}_i^Q,\boldsymbol{KW}_i^K,\boldsymbol{VW}_i^V$,在每一个空间中分别计算得到注意力向量

$$\boldsymbol{U}_i=\mathrm{softmax}\left(\frac{\boldsymbol{QW}_i^Q\cdot\boldsymbol{KW}_i^K}{\sqrt{d_k}}\right)\boldsymbol{VW}_i^V,$$

其中 d_k 是向量 $\boldsymbol{K}$ 的维度,防止通过点乘计算相似度时数值过大.多头注意力机制将所有空间的注意力向量进行拼接,并进行投影.

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\boldsymbol{U}_1,\cdots,\boldsymbol{U}_h)\boldsymbol{W}^o.$$

全连接层将多头注意力层的输出进行投影.全连接的形式为

$$\mathrm{FFN}(x)=\max(0,x\boldsymbol{W}_1+b_1)\boldsymbol{W}_2+b_2 \quad (1)$$

该全连接层在同一层的不同位置是共享的,不同层之间全连接层的结构是不同的.我们把 Transformer 第 i 层的输出记为 H_i.

我们把词向量表示输入到前文提到的 Transformer 的一层,在底层自我注意力机制中

$$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{E},$$

在随后的 11 层 Transformer 单元中

$$\boldsymbol{Q}_i=\boldsymbol{K}_i=\boldsymbol{V}_i=\boldsymbol{H}_{i-1}.$$

通过如上词嵌入层和 12 层 Transformer 单元,我们最终得到文本的表示 $H_1,\cdots,H_{12}$,我们利用第 12 层的输出作为文本的表示.注意 H_{12} 中包含问题和文本表示,我们不使用问题的表示部分,仅仅使用文本的表示.我们把文本表示部分记为 $\boldsymbol{H}^p$.

3.3 基于 RNN 的阅读模块

本文所提模型 BERT-Boundary 是采用 BERT 作为底层结构,基于答案边界序列标注的 Boundary

模块作为答案预测层的组合. 在 BERT 以前的模型多采用 RNN 作为底层结构，为了验证我们的 Boundary 是阅读模块无关的且一致有效，我们下边介绍一种基于 RNN 的阅读模块作为对比.

该部分阅读理解模型采用 LSTM 去编码信息，结合注意力机制去对齐文本和问题的内容，使得文本的表示能感知哪些内容是重要的，换言之就是和问题相关的，整体结构如图 3 所示.

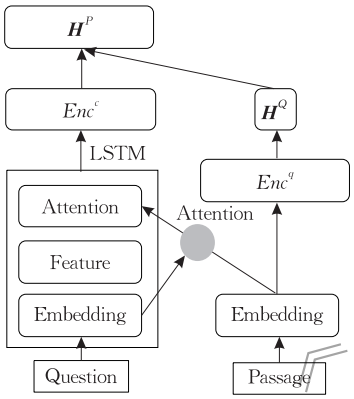

图 3　基于 RNN 的阅读模块结构

下面将整个模块分为三个部分进行介绍，分别为嵌入层、交互层和编码层.

嵌入层是将词序列转换成对应的向量，向量包含两部分内容，其一是词嵌入，通过嵌入层将词映射成词向量 $\boldsymbol{E}_{[p_i]}$，本文采用使用 300 维 GloVE 预训练的词向量，模型训练过程中词向量不更新，对于未出现在预训练词表中的词，本文采用随机初始化. 另一部分内容是语言学特征和统计特征，其中包括词性标注，命名实体识别和该词是否出现在问题中，以及词干还原后是否出现在问题中等，这些特征是一些二值特征或者是类别型特征. 将词 p_i 这些特征进行独热编码表示成向量 $\boldsymbol{F}_{[i]}$. 语言学特征和统计特征可以辅助和补充词的语义特征，帮助区分文本中重要的词片段，使用这些特征可以缓解模型学习难度，将本来需要模型隐式学习的内容直接显示提供给模型. 问题 q 中的词经过同样的词嵌入也将词转化为向量. 在文本的表示过程中引进问题的信息是必要的，我们通过词对齐即注意力机制可以完成这个目标. 本文使用文本中词嵌入向量 $\boldsymbol{E}_{[p_i]}$ 和问题的词嵌入序列 $E_{[q_1]},\cdots,E_{[q_m]}$ 中的每个向量计算一个相似度得分，然后通过该得分将问题的词嵌入进行加权求和，得到一个与该文本词嵌入 $\boldsymbol{E}_{[p_i]}$ 对齐的问题表示向量 $\boldsymbol{H}_{[i]}$，该向量表明文本词 p_i 和问题的一个相关程度. 具体公式如下：

$$a_{i,j}=\frac{\exp(\boldsymbol{E}_{[p_i]}\cdot E_{[q_j]})}{\sum_{j'}\exp(\boldsymbol{E}_{[p_i]}\cdot E_{[q_{j'}]})},$$

$$\boldsymbol{H}_{[i]}=\sum_j a_{i,j}E_{[q_j]}.$$

通过该对齐矩阵 $\boldsymbol{H}$，文本中的每个词丰富了其与问题的关联信息. 推理答案位置的过程本质上是在文本中匹配到与问题相似的部分，然后从周围找到答案片段. 已经对单个词项和问题之间的信息建模，本文需要进一步将这些信号收集.

在编码层，本文使用双向的 GRU 层进行上文信息收集. 在我们的实验中，LSTM 和 GRU 效果相当，但是 GRU 速度更快. 将 $\boldsymbol{E}_{[p_i]}$，$\boldsymbol{H}_{[i]}$ 和 $\boldsymbol{F}_{[i]}$ 拼接后输入到 GRU 层得到向量 $\boldsymbol{Enc}^P_{[i]}$，问题表示 $E_{[q_j]}$ 也同样输入 GRU 进行编码得到向量 $\boldsymbol{Enc}^Q_{[j]}$，公式如下：

$$\boldsymbol{Enc}^P_{[i]}=\mathrm{GRU}(\boldsymbol{Enc}^P_{[i-1]},[\boldsymbol{E}_{[p_i]};\boldsymbol{H}_{[i]};\boldsymbol{F}_{[i]}]),$$

$$\boldsymbol{Enc}^Q_{[j]}=\mathrm{GRU}(\boldsymbol{Enc}^Q_{[j-1]},E_{[q_j]}).$$

本文进一步将问题矩阵 $\boldsymbol{Enc}^Q$ 表示成一个包含问题整体语义的向量 $\boldsymbol{H}^Q$ 与文本表示 $\boldsymbol{Enc}^P$ 进行交互. 本文通过一个自注意力机制对原来问题矩阵进行加权求和. 公式如下：

$$\boldsymbol{H}^Q=\sum_j a_j\boldsymbol{Enc}^Q_{[j]},$$

$$a_j=\frac{\exp(\boldsymbol{w}\cdot\boldsymbol{Enc}^Q_{[j]})}{\sum_{j'}\exp(\boldsymbol{w}\cdot\boldsymbol{Enc}^Q_{[j']})}.$$

通过如上的加权操作，可以捕获问题的整体信息，权重 a_j 有帮于忽略掉其中不重要的词如停用词连接词等，保留问题中重要信息. 进一步我们将 $\boldsymbol{H}^Q$ 与 $\boldsymbol{Enc}^P_{[j]}$ 进行拼接得到最终的文本表示 $\boldsymbol{H}^p$.

3.4　Boundary：基于答案边界的答案预测模块

经过底层阅读理解模块（基于 BERT 或者基于 RNN），文本被表示为一个矩阵 $\boldsymbol{H}^p$，文本中的词项对应到一个向量的表示 $\boldsymbol{H}^p_i$. Boundary 答案预测模块对答案的边界进行建模，解决多片段答案的抽取，如图 4 所示. 通过两层全连接打分网络，为文本中每个词项计算两个分数，分数经过归一化形成概率. 每个词项相对应的两个概率值，具体公式如下：

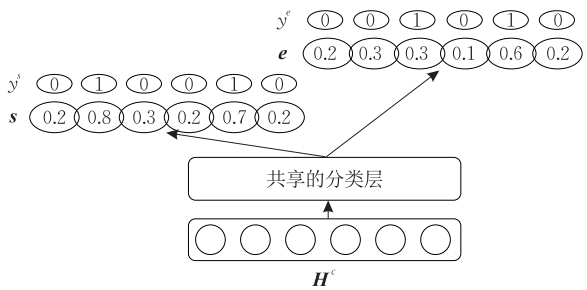

图 4　基于边界序列标注的 Boundary 答案预测模块

$$\boldsymbol{s}_i=\mathrm{sigmoid}(\mathrm{FFN}(\boldsymbol{H}^p_i)),$$

$$\boldsymbol{e}_i=\text{sigmoid}(\text{FFN}(\boldsymbol{H}_i^p)),$$

$\boldsymbol{s}_i$代表文本的第 i 个词作为答案片段起始的概率，$\boldsymbol{e}_i$表示文本的第 i 个词作为答案片段结束的概率，$\boldsymbol{s}$ 和 $\boldsymbol{e}$ 通过相同结构不同参数计算得出. 整个句子的表示得到两个分数向量 $\boldsymbol{s}$ 和 $\boldsymbol{e}$，其中 FFN 定义同式(1).

本文用的损失函数对模型产生的两个向量 $\boldsymbol{s}$ 和 $\boldsymbol{e}$ 的每个位置进行二分类，其中 $\boldsymbol{s}$ 拟合每个位置作为起始位置的概率，$\boldsymbol{e}$ 拟合每个位置作为结束位置的概率. 具体公式如下：

$$loss_s=\frac{1}{n}\sum_i -y_i^s\log(\boldsymbol{s}_i)-(1-y_i^s)\log(1-\boldsymbol{s}_i),$$

$$loss_e=\frac{1}{n}\sum_i -y_i^e\log(\boldsymbol{e}_i)-(1-y_i^e)\log(1-\boldsymbol{e}_i),$$

$$loss=loss_s+loss_e,$$

其中 y_i^s和 y_i^e是真实答案边界的指示，$y_i^s=1$ 表示第 i 词是答案的起始位置. 该方式克服以往阅读理解模型损失无法处理多片段答案问题，通过单独每个词进行分类其是否是起始位置，可以获得多个起始位置，同理可以获得多个结束位置. 不同于以前的模型直接预测起始位置分布和结束位置分布，从而导致其仅能建模一对起始结束位置.

3.5 解码算法

由于阅读理解模型需要输出的是答案片段，而模型的预测是两个分数向量 $\boldsymbol{s}$ 和 $\boldsymbol{e}$，解码算法用于将分数向量转化成最终的答案输出. 我们直接使用对数概率进行解码.

不同于单一答案片段时的解码算法，其选择概率最大片段，这样的算法依赖强的假设，即答案片段存在且唯一. 在多个答案片段的情形，该方法无法奏效，我们无法确定返回多少个片段. 因此本文提出顺序匹配的解码算法，该算法可以根据预测分数确定返回答案的数量，且保证返回答案不重叠. 具体如图 5 所示. 该算法顺序遍历起始分数向量和结束分数向量，把大于指定阈值索引记录在 cand_start 和 cand_end 数组中，然后遍历这两个数组将符合条件的片段进行输出. 注意该预测作为超参数，可以根据实际需要调节返回答案的准确率和召回率. 如果没有出现大于阈值的片段，就返回所有片段中起始加结束概率最大的片段.

3.6 Content：基于答案内容的答案预测模块

Li 等人[26]提出序列标注的方式去抽取答案. 本文将其迁移到多片段答案的抽取应用中，作为我们模型的一种基线方法. 该方法直接对答案内容进行标注，我们将其称为 Content 答案预测模块. 如图 6 所示，该模块基于阅读模型模型的文本表示 $\boldsymbol{H}^p$ 产生一个分数向量 $\boldsymbol{c}$.

$$\boldsymbol{c}_i=\text{Sigmoid}(\text{FFN}(\boldsymbol{H}_i^p)),$$

其中 $\boldsymbol{c}_i$表示第 i 个词作为答案内容的概率. 训练时分类每个词不是答案内容. 训练时采用损失函数形式如下：

$$loss=\frac{1}{n}\sum_i -y_i\log(\boldsymbol{c}_i)-(1-y_i)\log(1-\boldsymbol{c}_i),$$

其中 y_i代表文本中第 i 个词是否是答案中的词，如图 6 上边 01 序列所示，其取值为二值化的 0/1. 针对多片段答案情况，超过指定阈值的词被放入答案池，这些词所组成的多个片段作为最终的答案.

Algorithm 1 Answer decoding algorithm
Input: s, e, *thresh*
Output: *A* (answers in the array)

```
function DecodeAnswer(s, e)
    cand_start ← []
    cand_end ← []
    answers ← []
    for i ← 1 to s.length do
        if s[i] > thresh then
            cand_start.add(i)
        end if
    end for
    for i ← 1 to e.length do
        if e[i] > thresh then
            cand_start.add(i)
        end if
    end for
    for i ← cand_start do
        for j ← cand_end do
            if j ≤ i then
                answers.add((i, j))
            end if
        end for
    end for
    if answers.length = 0 then
        max_score = -1000
        max_span = ()
        for i ← 1 to s.length do
            for j ← 1 to e.length do
                if s[i] + e[j] > max_score then
                    answers.add((i, j))
                    max_span = s[i] + e[j]
                end if
            end for
        end for
    end if
    return answers
end function
```

图 5 顺序匹配的解码算法

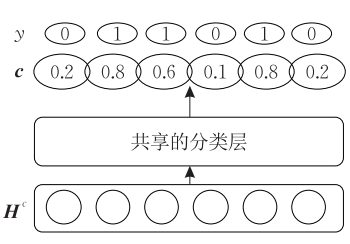

图 6 基于答案内容的 Content 答案预测模块

4　实验和分析

本节介绍实验和分析. 我们首先介绍基础实验设置：包括所用的数据和评价指标. 我们对两种阅读理解模块（基于 RNN 和基于 BERT），和两种答案预测模块（Boundary 和 Content）分别进行组合实验. 验证我们的模型 BERT-Boundary 是一个较优的设计. 同时我们对模型特性进行详细的分析.

4.1　实验设置

数据集. 本文构建了一个符合真实需求的面向多片段答案的阅读理解数据集合，该数据集合包含 153 297 个样本，从中划分出 10 000 条作为验证集，划分出 10 000 条作为测试集. 每条数据包括一段文本和一个问题，以及多个答案片段在文本中的位置. 数据的收集过程来自于商业搜索引擎的匿名用户查询，从中过滤出问题式查询，结合搜索引擎给出的相关文本段，人工从文本中标记出答案片段，要求标注出所有正确的答案片段且保持答案精简，不可以标注与答案上下文无关信息.

不同于以往的 SQuAD、TriviaQA 等数据集，答案片段仅由其中文本中的单独一个片段组成，本文所用数据集中的答案包含一个或者多个片段. 以上两个数据集的问题并不是来自真实的用户查询，本文的数据集合来自真实的用户查询.

和 DuReader 和 MS MARCO 相比较，其数据的查询同样来自真实的搜索引擎日志，答案为人工编辑，该类数据集可以反映真实用户的信息需求且形式多样，但是问题在于人工编辑的答案难以评价. 我们的多片段抽取式的阅读理解易于评价.

我们标注的数据包含问题，文本和多个答案片段. 例子如下所示.

例子 1.

文本：Salary. The average salary range for a zoologist in the initial stages of his or her career is **$30,000 to $45,000** per year. After five years of work experience, the range is **$40,000 to $55,000** per year.

问题：zoology salary

答案：$30,000 to $45,000, $40,000 to $55,000

最终收集的数据基本信息如表 1 所示，我们可以看到文本平均长度为 40 词，文本来自搜索引擎给出的最相关片段. 问题平均长度 6.01 词说明搜索引擎的查询一般较短. 长度短的查询一般其意图表达不明确，所以有可能出现多片段答案的情况. 如上边的例子所示，问题简短缺乏限定，文本中两处片段都是其答案. 每个答案平均由 1.76 个片段组成.

表 1　数据集基本信息统计

文本长度	问题长度	数据规模	答案片段数量
40.23	6.01	153 297	1.76

我们接下来详细分析答案片段数量分布，我们统计由不同数量片段组成的答案的所占比例如图 7 所示.

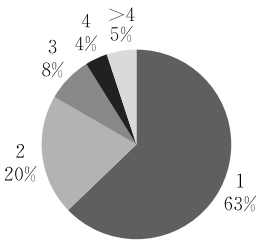

图 7　不同数目答案片段的组成的答案的样本分布

从图中可以看到答案由多个片段组成的情况占比高达 37%，其中有 20% 由两个答案片段组成，17% 由多于两个答案片段组成. 这说明在实际搜索场景中，多个片段的情况较为常见. 我们统计答案对应的答案片段数量为 1～5 时，对应的平均文本长度如表 2 所示. 从中可以看出，搜索中的简短模糊的查询是造成答案为片段的一个原因.

表 2　问题长度和答案片段数量统计

答案片段数量	平均文本长度
1	6.11
2	5.93
3	5.67
4	5.57
5	5.54

一个可能的想法是是否可以通过扩张单个片段的长度把多个片段都涵盖到一个片段中，比如在例子 1 中我们找到包含“$30,000 to $45,000”和“$40,000 to $55,000”最小片段，我们可以看到其内容基本弥散在整个文本中，包含两个片段的最小文本片段包含大量冗余信息，不符合阅读理解的初衷，找出能回答问题的精确的片段.

评价指标. 实验评价指标采用衡量预测出的答案片段和真实的答案片段是否一致的完全匹配（下文简称为 *EM*），和部分一致的片段级别 *F*1 指标（下文简称 *F*1）.

预测的答案片段集合为 $\boldsymbol{A}^p=\{A_1^p,\cdots,A_{k'}^p\}$，真

实的答案片段集合为$\boldsymbol{A}^{gt}=\{A_1^{gt},\cdots,A_k^{gt}\}$.对于该条样本 EM 值为 1 当且仅当 $\boldsymbol{A}^{p}$ 和 $\boldsymbol{A}^{gt}$ 完全相等,即数量和每个片段内容都一致.

答案片段级别的 $F1$ 指标衡量的是两者部分匹配程度,为了给出 $F1$ 指标,本文首先给出片段级别的精确率(*precision*)和召回率(*recall*)定义:

$$precision=\frac{|\boldsymbol{A}^{p}\bigcap\boldsymbol{A}^{gt}|}{|\boldsymbol{A}^{p}|},$$

$$recall=\frac{|\boldsymbol{A}^{p}\bigcap\boldsymbol{A}^{gt}|}{|\boldsymbol{A}^{gt}|},$$

$$F1=\frac{2\times precision\times recall}{precision+recall},$$

其中 *precision* 指示预测答案片段有多少比例是正确的,*recall* 指示被正确抽取出的正确答案的比例.$F1$ 值则是 *precision* 和 *recall* 的调和值.

4.2 对比模型

本文对比模型主要有以往阅读理解模型和一些变种.以往的阅读理解模型包括 BIDAF 和 BERT,这两个模型分别是基于 RNN 和基于预训练的有代表性的模型.另外我们提出基于 Li[26] 改进的基于内容标注的 Content 模块和 RNN、BERT 阅读模块进行结合产生的 RNN-Content 模型和 BERT-Content 模型,同时我们把我们所提的 Boundary 答案预测模块和 RNN 阅读模块结合为 RNN-Boundary.

对于 BIDAF 模型和 BERT 模型的训练,它们无法学习多个答案片段的信息,本文从多片段答案随机选择一个片段作为其学习目标,训练和预测均采用其官方提供的源码.

对于本文改进 RNN-Content 模型和 BERT-Content 模型,其可以学习多个答案片段位置.本文把数据集的位置生成 n 维向量,n 是相关文本的长度,向量的元素为 1 当且仅当该元素位置处于真实答案片段中.

对于 RNN-Boundary 和 BERT-Boundary 模型,训练数据的位置信息转化两个长度为 n 的向量,第一个向量的元素指示该位置是否是答案片段的起始位置,另一个向量的元素指示该位置是否是答案片段的结束位置.

三类模型学习的标签示例如下:

Text:中国 的 一线 城市 有 ***北京*** 和 ***上海*** 两个 直辖市,以及 广东省 两个 工业 重镇 ***广州*** 和 ***深圳***

BIDAF 和 BERT 的学习标签(其中之一):

[5,5],[7,7],[16,16][18,18]

RNN-Content 和 BERT-Content 学习的标签:

[0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,1,0,1]

RNN-Boundary 和 BERT-Boundary 学习的标签:

[0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,1,0,1]

[0,0,0,0,0,1,0,1,0,0,0,0,0,0,0,1,0,1]

4.3 实验结果

本文使用训练集数据训练各个模型,在验证集上根据 $F1$ 指标去确定模型的最优训练轮数,几种方法在测试集合的性能表现如图 8 所示.从图中我们可以看出 BERT 作为阅读模块的性能要好于基于 RNN 的方法.我们再对比原始阅读理解模型和加上 Content,Boundary 答案预测模块后的效果,我们可以看到原有阅读理解模型不能建模多个单位置,导致其 EM 和 $F1$ 指标均为最低.我们提出 Boundary 多片段建模方法要好于 Content 答案预测模块,且在基于 RNN 和基于 BERT 两种阅读理解模块上 Boundary 一致好于 Content 模块.最后我们可以看到我们的模型 BERT-Boundary 取得最好效果,$F1$ 指标上较 BERT-Content 提升 2.5%,较 BERT 提升 18%.

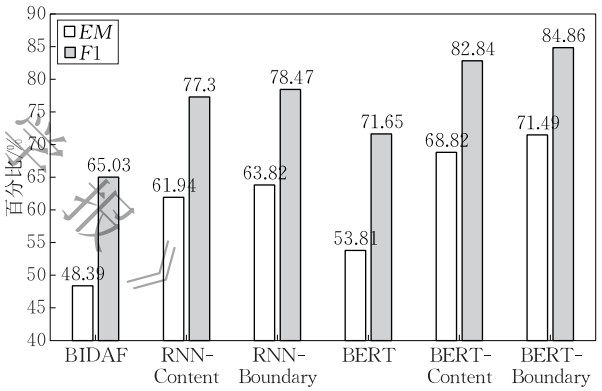

图 8 各个方法在测试集的实验结果

4.4 分析实验

我们把训练集和测试集分为单片段部分和多片段部分,在相应部分上训练和预测.测试 BERT,BERT-Content 和我们的 BERT-Boundary 模型的效果.

从表 3 在测试集单片段上的实验结果中可以看出在单片段上我们模型 BERT-Boundary 略弱于原始 BERT,这是由于我们的模型去除了答案仅由单片段组成的假设,模型有预测出多个片段的风险,所以我们的 EM 和 $F1$ 值略低.但是对比 BERT-Content 可以看出,我们模型在单片段的的损失要小于 BERT-Content.可能的原因是 Content 答案预测模块解码困难.

表 3　在测试集单片段上的实验结果

Model	*EM*	*F*1
BERT	88.31	88.31
BERT-Content	85.89	86.40
BERT-Boundary	87.00	87.31

从表 4 多片段结果我们可以看出原有模型无法预测多片段，其 *EM* 始终为 0，*F*1 值很低，由于其最多能正确预测出其中一个片段. BERT-Boundary 性能也远超过 BERT-Content 模型.

表 4　在测试集多片段部分的实验结果

Model	*EM*	*F*1
BERT	0	47.00
BERT-Content	56.02	80.65
BERT-Boundary	59.57	85.17

我们在 SQuAD 数据集合上测试 BERT 和 BERT-Boundary 效果，可以看出我们的多片段阅读理解较原来 BERT 在单片段数据有轻微下降. 这是去除单片段的限制的代价，可以预测不定数量的答案片段，也导致在单片段数据集合上可能错误地预测出多个片段. 如果我们把单个答案片段假设在解码阶段加入到 BERT-Boundary，即解码时只取分数最高的答案，我们的模型可以达到和 BERT 一样的性能.

我们进一步分析各个模型在不同答案片段数量上的表现，本文按答案片段数量将测试集合分成 5 个部分，其分别为答案片段数量为 1 个到 5 个. 本文绘制如图 9、图 10 的 *EM* 和 *F*1 指标的随着答案片段数量的折线图. 从表 5 中我们可以看到，当答案

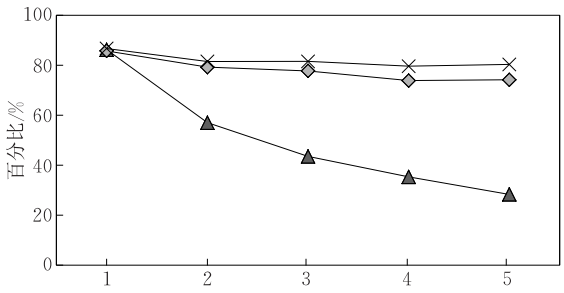

图 9　模型 *F*1 随答案片段数量的变化

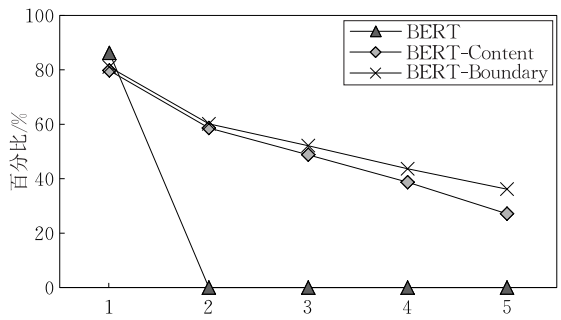

图 10　模型 *EM* 值随答案片段数量的变化

表 5　SQuAD 数据集上的结果

Model	*EM*	*F*1
BERT	79.176	87.590
BERT-Boundary	78.174	86.762

片段数量变多，准确地抽取出所有片段变得困难，所有模型性能都有所下降，但是我们的 BERT-Boundary 下降幅度缓慢. 能够在 5 个片段时依然维持 *F*1 值在 80 以上的性能.

从图 10 中我们可以看到 BERT 在片段数量大于 1 时 *EM* 指标直接降到 0，由于其无法预测多个答案，我们尝试根据片段概率直接从原有 BERT 解码多个答案，它们之间相互重叠，仍然不能提高其 *EM* 值. BERT-Boundary 较 BERT-Content 性能下降略缓. 通过图 9 和图 10 中 BERT-Boundary 曲线我们可以发现，文本片段数量增加时，模型同时预测正确的难度急剧增加. *EM* 指标下降速度较快.

本文将测试数据集合按照答案片段长度分成 5 个区间，计算模型在每个答案长度区间上的表现. 对于多个片段答案的问题，对各个片段长度求平均作为答案片段长度. 绘制如图 11 的 *F*1 性能随着答案片段长度的变化直方图.

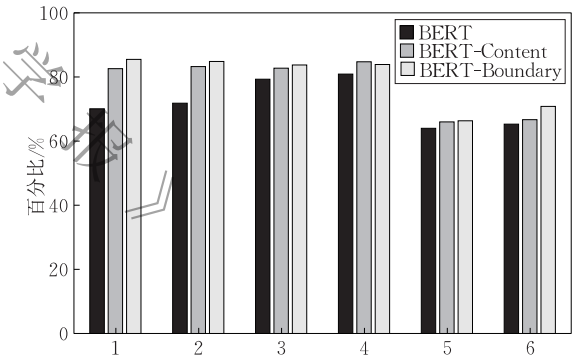

图 11　模型 *F*1 随答案片段数量的变化

从中可以看出本文的方法在各个答案片段长度上表现差距不大，在答案长度达到 5 以上时出现一定下降. BERT-Boundary 在长度较长时好于 BBERT-Content 方法. 本文的方法直接对起始位置和结束位置分类，对起始位置和结束位置的分类可以很好的利用答案片段周围的上下文信息，而 Content 模块当答案片段较长时，其单个答案片段中间的词由于距离上下文较远可能会出现预测错误，这样的预测错误会导致预测的答案片段不完整. 另外我们看到 BERT 性能在长度为 3、4 时最佳，我们分析后得知由于文本 BERT 不能预测多片段，长度为 3、4 片段多出现在单片答案中，所以此时 BERT 表现略好.

此外本文的方法可以通过调节阈值去控制返回

答案的多少，调高阈值答案数量会变少，调低阈值返回的答案数量会增加. 当把模型应用到搜索引擎的高亮中，把跟答案相关的部分高亮显示，此时更加注重正确答案片段的召回率，可以调低阈值使得召回率增加.

5 总结和展望

本文提出一种新的多片段答案的阅读理解模型 BERT-Boundary，该模型利用 BERT 结构作为阅读模块进行文本和问题的理解，利用 Boundary 作为答案预测模块建模多个答案片段. 通过和已有阅读理解模型以及改进的朴素多片段解决方案通过实验对比，我们方法在多片段阅读理解数据集合上取得一致的提升. 本文详细分析模型在不同维度上的表现，验证了其结果的有效性.

目前该方法还是单独去建模一个文本中的多个答案片段，这些答案片段之间本质上是相互关联的，未来的研究工作可以围绕多个片段之间的关联展开，将答案片段之间的关联考虑进去，进一步提升多个片段答案预测的准确度.

参 考 文 献

[1] Rajpurkar P, Zhang J, Lopyrev K, et al. SQuAD: 100 000+ Questions for machine comprehension of text//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Texas, USA, 2016: 2383-2392

[2] Joshi M, Choi E, Weld D, et al. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 1601-1611

[3] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[4] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend//Proceedings of the Advances in Neural Information Processing Systems. Montréal, Canada, 2015: 1693-1701

[5] Hill F, Bordes A, Chopra S, et al. The goldilocks principle: Reading children's books with explicit memory representations //Proceedings of the ICLR 2015. San Diego, USA, 2015

[6] Richardson M, Burges C J C, Renshaw E. MCTest: A challenge dataset for the open-domain machine comprehension of text// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 193-203

[7] Lai G, Xie Q, Liu H, et al. RACE: Large-scale reading comprehension dataset from examinations//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 785-794

[8] Talmor A, Herzig J, Lourie N, et al. CommonsenseQA: A question answering challenge targeting commonsense knowledge //Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4149-4158

[9] Yagcioglu S, Erdem A, Erdem E, et al. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1358-1368

[10] Dunn M, Sagun L, Higgins M, et al. SearchQA: A new Q&A dataset augmented with context from a search engine. arXiv preprint arXiv:1704.05179, 2017

[11] He W, Liu K, Liu J, et al. DuReader: A Chinese machine reading comprehension dataset from real-world applications// Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne, Australia, 2018: 37-46

[12] Nguyen T, Rosenberg M, Song X, et al. MS MARCO: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268, 2016

[13] Reddy S, Chen D, Manning C D. CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 2019, 7: 249-266

[14] Choi E, He H, Iyyer M, et al. QuAC: Question answering in context//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 2174-2184

[15] Wang S, Jiang J. Machine comprehension using match-LSTM and answer pointer//Proceedings of the ICLR 2017. Toulon, France, 2017

[16] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension//Proceedings of the ICLR 2017. Toulon, France, 2017

[17] Wang W, Yang N, Wei F, et al. Gated self-matching networks for reading comprehension and question answering //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 189-198

[18] Huang H Y, Zhu C, Shen Y, et al. FusionNet: Fusing via fully-aware attention with application to machine comprehension //Proceedings of the ICLR 2018. Vancouver Canada, 2018

[19] Wang W, Yan M, Wu C. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 1705-1714

[20] McCann B, Bradbury J, Xiong C, et al. Learned in translation: Contextualized word vectors//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 6294-6305

[21] Peters M, Neumann M, Iyyer M, et al. Deep contextualized word representations//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Louisiana, USA, 2018: 2227-2237

[22] Hu M, Peng Y, Huang Z, et al. Reinforced mnemonic reader for machine reading comprehension//Proceedings of the 27th International Joint Conference on Artificial Intelligence. LA, USA, 2018: 4099-4106

[23] Liu X, Shen Y, Duh K, et al. Stochastic answer networks for machine reading comprehension//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 1694-1704

[24] Chen D, Bolton J, Manning C D. A thorough examination of the CNN/daily mail reading comprehension task//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). BerlinGermany, 2016: 2358-2367

[25] Wang S, Yu M, Jiang J, et al. A co-matching model for multi-choice reading comprehension//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia, 2018: 746-751

[26] Li P, Li W, He Z, et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. arXiv preprint arXiv:1607.06275, 2016

[27] Wang S, Yu M, Guo X, et al. R^3: Reinforced ranker-reader for open-domain question answering//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. LA, USA, 2018: 5981-5988

[28] Lin Y, Ji H, Liu Z, et al. Denoising distantly supervised open-domain question answering//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 1736-1745

[29] Wang Y, Liu K, Liu J, et al. Multi-passage machine reading comprehension with cross-passage answer verification//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 1918-1927

[30] Yan M, Xia J, Wu C, et al. A deep cascade model for multi-document reading comprehension//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 7354-7361

SU Li-Xin, Ph. D. candidate. His research interests include information retrieval and automatic question answering.

GUO Jia-Feng, Ph. D., professor. His research interests include information retrieval and data mining.

FAN Yi-Xing, Ph. D., assistant professor. His research interests include information retrieval and data mining.

LAN Yan-Yan, Ph. D., associate professor. Her research interests include statistics machine learning, learning to rank and information retrieval.

XU Jun, Ph. D., professor. His research interests include information retrieval and data mining.

CHENG Xue-Qi, Ph. D., professor. His research interests include network science, network and information security, web search and data mining, etc.

Background

This work addresses the multi-span answer reading comprehension. This is newly proposed task in NLP community.

The multi-span answer reading comprehension is critical for applying the reading comprehension model in real system, as there are many questions or queries whose answers are composed of multiple span from the context passages.

In this paper, we proposed BERT-Boundary model which uses the pre-trained BERT as lower layer and employs an answer-boundary labeling model for locating multiple answer spans. BERT-boundary combines the power of text understanding and multi-span locating. We conduct extensive experiments on the newly constructed multi-span answer dataset. The experimental results demonstrate the effectiveness of our model and a clear improvement on $F1$ measurement over existing models.

The authors of the paper have done lots of research on information retrieval and natural language processing, like question answering. They proposed a risk control framework in question answering in SIGIR 2019 conference.

This work was funded by the National Natural Science Foundation of China under Grant Nos. 61425016, 61722211, 61773362, and 61872338, the Youth Innovation Promotion Association CAS under Grants Nos. 20144310, and 2016102, the National Key R&D Program of China under Grant No. 2016QY02D0405, and the Foundation and Frontier Research Key Program of Chongqing Science and Technology Commission (No. cstc2017jcyjBX0059).