CCL23-Eval 任务9系统报告：基于重叠片段生成增强阅读理解模型鲁棒性的方法

何苏哲
北京理工大学
北京市海量语言信息处理与云计算应用工程技术研究中心
hesuzhe@163.com

杨崇盛
北京理工大学
北京市海量语言信息处理与云计算应用工程技术研究中心
csyang@bit.edu.cn

史树敏*
北京理工大学
北京市海量语言信息处理与云计算应用工程技术研究中心
bjssm@bit.edu.cn

摘要

目前机器阅读理解在抽取语义完整的选项证据时存在诸多挑战。现有通过无监督方式进行证据抽取的工作主要分为两类，一是利用静态词向量，采用集束搜索迭代地提取相关句子；另一类是使用实例级监督方法，包括独立式证据抽取和端到端式证据抽取。前者处理流程上较为繁琐，后者在联合训练时存在不稳定性，直接导致模型性能难以稳定提升。在CCL23-Eval 任务9中，本文提出了一种基于重叠片段生成的自适应端到端证据抽取方法。该方法针对证据句边界不明确的问题，通过将文档划分为多个重叠的句子片段，并提取关键部分作为证据来实现整体语义的抽取。同时，将证据提取嵌入模块予以优化，实现了证据片段置信度自动调整。实验结果表明本文所提出方法能够极大地排除冗余内容干扰，仅需一个超参数即可稳定提升阅读理解模型性能，增强了模型鲁棒性。

关键词： 重叠片段生成；无监督证据抽取；机器阅读理解；鲁棒性；置信度

System Report for CCL23-Eval Task 9: Improving MRC Robustness with Overlapping Segments Generation for GCRC_advRobust

Suzhe He
Beijing Institute of Technology Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
hesuzhe@163.com

Chongsheng Yang
Beijing Institute of Technology Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
csyang@bit.edu.cn

Shumin Shi*
Beijing Institute of Technology Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
bjssm@bit.edu.cn

Abstract

There are many challenges in machine reading comprehension when it comes to extracting semantically complete evidence for specific statement. Existing works on unsupervised evidence extraction can be mainly divided into two categories. The first category

* Corresponding Author

utilizes static word vectors and employs beam search iteratively to extract relevant sentences. The second category uses instance-level supervised methods, including independent evidence extraction and end-to-end evidence extraction. The former category involves a more complex process, while the latter category suffers from instability during joint training, which directly affects the stable improvement of model performance. In Task 9 of CCL23-Eval, this paper proposes an adaptive end-to-end evidence extraction method based on overlapping segments generation. This method addresses the uncertainty in evidence sentence boundaries by dividing the document into multiple overlapping sentence segments and extracting key parts as evidence to achieve overall semantic extraction. Simultaneously, the evidence extraction embedding module is optimized to achieve automatic adjustment of evidence segment confidence. Experimental results demonstrate that the proposed method greatly eliminates interference from redundant content, and with just one hyperparameter, it can stably improve the performance of the reading comprehension model, improving model robustness.

Keywords: Overlapping Segments Generation , Unsupervised Evidence Extraction , Machine Reading Comprehension , Robustness , Confidence

1 引言

近年来，随着预训练模型的出现和大规模数据的提出，自然语言处理中的任务性能得到了大幅提升，甚至有些超过人类。但是没有理由的预测适用性有限，一些任务需要计算机在进行答案预测的同时进行答案相关证据进行提取，这就需要有证据标注的数据以供模型学习。比如，对于机器阅读理解，单纯的文章分块训练可能存在适用性的限制，因为这些任务需要模型理解文章，找到相关证据后进行答案预测。证据通常表现为篇章中的一个或多个连续序列，序列一般以句为单位。Figure 1为VGaokao 数据集(Zhang et al., 2021)的样例：

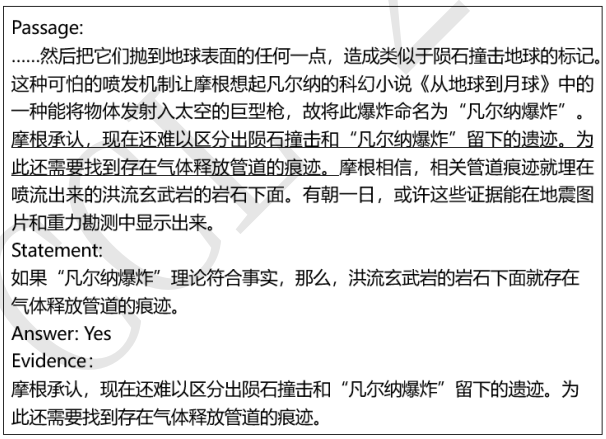

Passage:
.....然后把它们抛到地球表面的任何一点，造成类似于陨石撞击地球的标记。这种可怕的喷发机制让摩根想起凡尔纳的科幻小说《从地球到月球》中的一种能将物体发射入太空的巨型枪，故将此爆炸命名为“凡尔纳爆炸”。摩根承认，现在还难以区分出陨石撞击和“凡尔纳爆炸”留下的遗迹。为此还需要找到存在气体释放管道的痕迹。摩根相信，相关管道痕迹就埋在喷流出来的洪流玄武岩的岩石下面。有朝一日，或许这些证据能在地震图片和重力勘测中显示出来。
Statement:
如果“凡尔纳爆炸”理论符合事实，那么，洪流玄武岩的岩石下面就存在气体释放管道的痕迹。
Answer: Yes
Evidence:
摩根承认，现在还难以区分出陨石撞击和“凡尔纳爆炸”留下的遗迹。为此还需要找到存在气体释放管道的痕迹。

Figure 1: VGaokao 数据集的样例

对于论述“如果“凡尔纳爆炸”理论符合事实，那么，洪流玄武岩的岩石下面就存在气体释放管道的痕迹。”模型不仅需要回答文档是否支持此论述，还需要抽取出支持论述的证据“摩根承认，现在还难以区分出陨石撞击和“凡尔纳爆炸”留下的遗迹。为此还需要找到存在气体释放管道的痕迹。”其中证据包含两句话，抽取符合人类理性的证据需要保留整体语义。特别是在中文中句子难以界定(Wang et al., 2021)的情况下，如何抽取语义完整的证据是一个挑战。

本文提出了一种基于自适应的端到端证据抽取方法。对于如何抽取出具有整体语义的问题，本文提出将文档以句子为单位划分多个片段，对关键部分进行提取作为证据，对于端到端范式需要多个超参数导致训练不稳定的问题，本文提出将证据提取模块嵌入原模型一同进行优化，仅需要唯一超参数的条件下自动调整证据片段置信度，而且在重读排除干扰的关键内容（证据）后，能够稳定提升模型性能。

2 相关工作

早期机器阅读理解研究侧重于建模问题和参考文档之间的语义匹配。之后为了模拟人类的阅读模式，提出了从粗到细的分级方法。这样的模型首先阅读全文以选择相关的文本跨度，然后从这些相关跨度中推断答案。Li et al. (2018)提出了一种类似人类的阅读策略，该策略与学生进行阅读理解测试时的逻辑相似并结合了对文档和问题的一般理解。在长文档摘要场景下突出部分对模型复杂度降低和性能提升有着重要作用，Bajaj et al. (2021)尝试通过使用基于GPT-2语言模型困惑评分的新算法，在低资源机制下运行，通过识别源中最能支撑摘要的突出句子，压缩这些长文档，在实验中还发现识别出的突出句子往往与领域专家的独立人类标签一致。

由于人工标注证据代价高昂，之前的工作仅通过无监督的方式进行证据抽取，第一种无监督方法借助静态词向量，一些工作工作使用集束搜索（Beam Search）的方法迭代提取相关句，Li and Gaussier (2021)首先通过本地查询块预排序来选择长文档的关键重读块，然后聚合几个块以形成一个短文档，该短文档可以由BERT 等模型处理，有些工作改进了Beam Search方法。然而，这种无监督方法不能模拟人类推理过程进而理解文本。

而另一种无监督方法使用实例级监督，常用范式包括独立式证据抽取和端到端证据抽取方法，独立式提取由于多阶段特性其流程较为繁琐，而端到端提取方法将问题分成两个子任务解决，分别是提取和预测，提取任务通过构造提供忠实的解释，从输入中离散地提取片段，并传递给预测器，预测器利用这些片段做出预测，然后将两个任务联合进行训练优化。但是这种端到端的方法训练非常不稳定，无法有效转移到其他数据其它任务和帮助提升模型性能。

一方面端到端证据抽取方法抽取是离散的，是对词级别进行抽取，不能形成具有完整语义的证据进而形成弱监督数据供模型学习。另一方面端到端方法涉及多个超参数导致了结果不稳定，由于需要对提取和预测两个模型进行联合训练，在提取时需要对生成的隐向量$\mathbf{z}$ 进行约束，旨在规范证据抽取的简洁性和连续性。具有代表性的Lei et al. (2016)提出的模型损失函数如式1所示：

$$Loss_{total} = Loss\left(\mathbf{z}, x, y\right) + \lambda_1 \left\|\mathbf{z}\right\| + \lambda_2 \sum_t abs\left(\mathbf{z_t} - \mathbf{z_{t-1}}\right) \quad \text{(式1)}$$

其中$Loss(\mathbf{z}, x, y)$ 为输入x 经过提取后输入预测器的预测，与真实标签之间的标准损失，可以使用常见的分类损失来实现，例如交叉熵损失，表示提取的证据必须足以替代输入文本。为了规范抽取的证据，需要对长度为输入序列长度的隐向量$\mathbf{z}$ 进行约束，式1中第二项和第三项分别表示惩罚选择词的数量和不鼓励转换（鼓励选择的连续性）。由于需要多个超参数（λ_1和λ_2），导致结果相当不稳定，不能迁移到其他任务和数据，最后因为联合优化是不可微的，所以一般使用强化式估计对模型进行端到端训练(Williams, 1992)。

为了解决端到端方法的局限性和受识别和重读突出部分后能够帮助模型理解文本语义的启发，本文提出将证据抽取模块嵌入原模型，在共同优化调整的情况下使模型充分利用原文信息，能够在稳定提升模型性能同时抽取突出部分作为证据片段。

3 模型框架与方法

本节从模型整体框架（如Figure 2所示）出发，介绍自适应的端到端证据抽取的“整体与部分”训练流程，首先介绍引入整体语义的模型结构，然后对证据片段（部分）抽取模块进行详细描述，最后说明如何添加到整体模型中实现重读关键部分。

3.1 整体语义引入

本文面向多选阅读理解场景提出一种基于自适应的端到端证据抽取方法，该方法结构可以分为两部分：整体与部分。首先整体旨在整体语义的融入，让模型在全局信息中粗读文本并且做出预测。这部分的模型结构如Figure 2左半部分所示。

首先将陈述O（在需要问题信息时表现为问题与选项的拼接）和文档D 按照“[CLS]+ 陈述O+[SEP]+ 文档D+[SEP]”的格式进行拼接后输入模型M，其中[CLS] 和[SEP] 是用于分隔不同部分的特殊标记，模型M 由预训练模型和一层全连接层组成，模型M 中使用[CLS] 特殊标记的向量表示输入全连接层，输出为每个类别的概率值，经过$softmax$ 函数后，训练的损失函数为交叉熵损失：

第二十二届中国计算语言学大会论文集，第293页-第302页，哈尔滨，中国，2023年8月3日至5日。

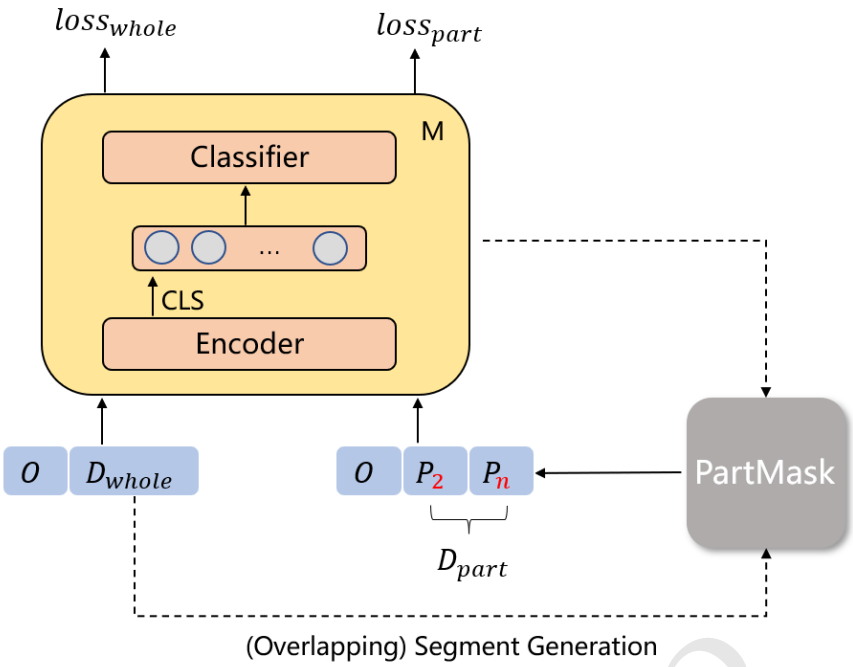


Figure 2: 整体模型结构

$$L_{total} = -log(P_y) \tag{式2}$$

3.2 非证据片段遮蔽

Figure 3展示了非证据片段遮蔽模块，此步骤分为文章片段生成、证据片段识别和证据片段抽取三部分，分别对应图中序号①、②和③，首先，为了抽取出语义完整的证据，对文档进行相对合理的片段划分，具体先将文档D 按句划分成t 句话，然后对其按式3进行部分（片段）的划分。

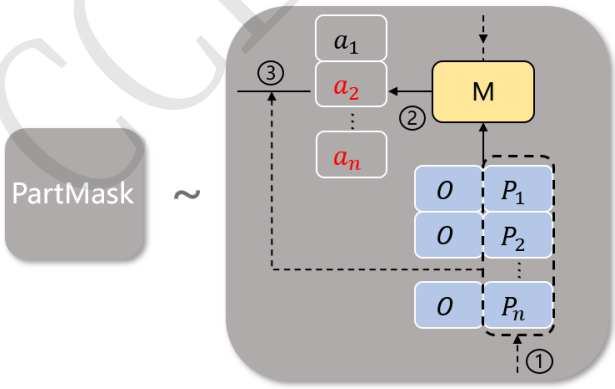


Figure 3: 证据片段抽取模块

$$n = \lceil t/r \rceil \tag{式3}$$

其中，

$$r = \begin{cases} 1 & if \quad t \leq k \\ \lceil t/k \rceil & if \quad t > k \end{cases}$$

第二十二届中国计算语言学大会论文集，第293页-第302页，哈尔滨，中国，2023年8月3日至5日。

其中r 为需合并的句子数，k 为文章依照句子划分出部份数的上限。使用向上取整保证了部分数$n \leq k$，n 为需最终划分的片段数。

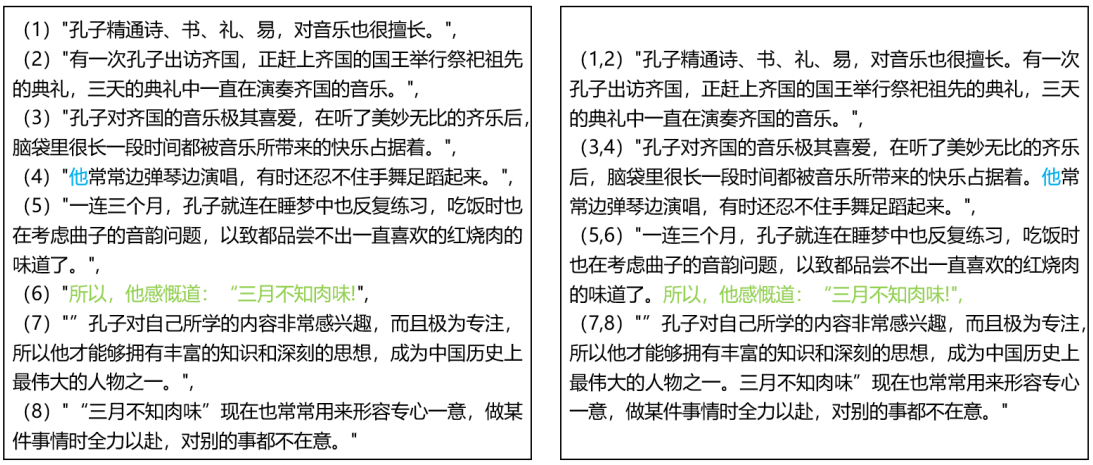


Figure 4: 文档划分样例

如Figure 4将文档（a）划分为（b）中多个部分，在这种灵活部分划分下，能够使每个部分语义更加完整，如原文（a）的带颜色标记的主语“他”表示“孔子”，只按句划分指向无法明确，而划分后融入了之前的语句，使每个部分都具有完整语境。最终将文档D 分为n 个片段：$D = \{P_1, P_2, ..., P_n\}$。以上是完整的无重叠片段生成操作细节。

接下来介绍证据片段识别，其目的是获得每个划分片段的置信度以供排序后抽取，首先将每个部分P_i 替换文档D，同样按照“[CLS]+ 陈述O +[SEP]+ 部分P_i+[SEP]” 的格式进行拼接输入模型M，将输出的概率值表示为每个部分作为证据的置信度v。由于每个样本的证据片段数量不同，需要对置信度v 进行标准化实现动态证据抽取，为了进一步扩大置信分数之间的差异，运用如式4所示的$Gumbel - Softmax$ 重参数化方法：

$$a_i = \frac{exp((log(v_i)+g_i)/T)}{\sum_j exp((log(v_j)+g_j)/T)} \quad i \in [1, n] \tag{式4}$$

其中T 是温度系数，为$Gumbel-Softmax$ 函数的超参数。为每个部分置信分数标准化后，可以对关键部分进行动态抽取，这里将置信分数a_i 大于β 的部分作为关键部分。在Figure 3右半部分例子中，从n 个片段$\{P_1, P_2, ..., P_n\}$ 中抽取出了P_2 和P_n 两个关键部分，将关键部分按原文档中顺序排列后形成新文档D_{part} 。证据抽取模块在无监督数据情况下不参与模型训练，用于关键部分抽取后形成新文档间接的参与训练，而在有监督情况下可以对置信分数实现损失计算，已达到更好的效果。

3.3 证据片段重读

为了精读证据片段，在生成新文档D_{part} 后，按照“[CLS]+ 陈述O+[SEP]+ 新文档D_{part}+[SEP]”的格式拼接后输入模型M，训练的损失函数为分类的交叉熵损失。最终的损失函数设计如式5所示：

$$Loss_{total} = Loss_{whole} + \alpha Loss_{part} \quad \alpha \in [0, 1] \tag{式5}$$

3.4 整体算法

本文方法整体算法流程如Algorithm 1所示，训练期间使用总损失函数$Loss_{total}$ 进行模型优化，为了衡量重读模块对模型性能的提升，只使用$Loss_{whole}$ 部分得到的概率值（即Figure 3左半部分）进行预测。最后保存所有训练周期的最优模型，重新获取3.2小节介绍的文档每个片段的置信分数a_i ，同样将置信分数a_i 大于β 的片段抽取为证据片段。

Algorithm 1 Training Algorithm

Input: Initial model M, Document D_{total}, Statement O(or concatenation of question and option)
Output: Trained model M, Evidence segments set D_{part}
Initialize model M randomly
Get $\{P_1, P_2, ..., P_n\} = D_{total}$ using document segments generation // Figure 3 step ①
for All epochs **do**
  Calculate $Loss_{total}$ according to M, D_{total}, O
  Get $D_{part} = \{P_i, P_j, ...\}$ using evidence segment identification // Figure 3 step ②
  Calculate $Loss_{part}$ according to M, D_{part}, O
  Train M by combining $Loss_{total}$ and $Loss_{part}$
end for
Get D_{part} according to M **return** M, D_{part}

以上，模型完成了整个训练与预测过程。首先进行整体信息引入，让模型获得全局词向量表示，之后对原文档关键部分抽取形成新文档，嵌入模型中重读关键信息，以求优化模型性能和准确抽取证据片段。

4 实验

4.1 数据集和评价指标

GCRC_advRobust(Tan et al., 2021)数据集是竞赛数据集，为衡量机器阅读理解模型的鲁棒性而设计，其训练集、验证集和测试集题目数分别为6994、336 和288，每题包括四个选项，GCRC 包含关键词扰动、推理逻辑扰动、时空属性扰动和因果关系扰动四种对抗攻击策略，每个待测试样本都有正负对抗选项，其具体划分如Table 1所示。

数据集划分	验证集	测试集
问题/侯选项数量	336/1344	288/1152
关键字词扰动侯选项数量	504	418
推理逻辑扰动侯选项数量	619	543
因果关系扰动侯选项数量	192	172
时空属性扰动侯选项数量	29	19

Table 1: GCRC_advRobust 数据集

系统的最终得分由Acc_0、Acc_1、Acc_2三个指标加权求和决定，具体计算式为：Score=0.2*Acc_0+0.3*Acc_1+0.5*Acc_2。其中，Acc_0为原始题目正确预测个数/题目总数，Acc_1为原始题目和任意一个对抗题目正确预测个数/题目总数，Acc_2为原始选项和两个对抗题目均正确预测个数/题目总数。

4.2 实验设置

实验均在一台NVIDIA GeForce RTX 3090 上运行。在证据抽取模块中将k 设为4，对于$Gumbel - Softmax$ 函数的超参数T 固定为0.5，对于决定抽取证据片段数的参数β ，由于所需证据片段数不同，为GCRC_advRobust 将β分别设为0.4。所有实验均在相同的环境中进行。训练批次设置为32，学习率为3.0e-5，训练周期设为10，编码器使用MacBERT-base预训练模型。对于损失函数，实验记录了$\alpha \in \{0.1, 0.5, 1\}$ 的最优准确率。

对GCRC_advRobust 数据集进行分块训练，测试时使用分块预测策略。实验中分析比较本文提出的添加关键重读模块（**E**vidence **F**ragment **R**ereading）的方法和基线模型的效果，并且侧面验证自适应证据抽取方法的有效性。基线模型不增加关键重读模块，直接在数据集上进行微调，如Figure 2中的左半部分。

基线模型（即baseline）将编码后的序列输入一个全连接层，从而得到多分类的概率值。EFR 为本文所提出的方法，在整体语义上进训练的同时，也对文档关键部分进行重读，最后经过自适应训练后可对文档进行证据片段抽取并衡量其有效性。

4.3 实验结果

模型方法	GCRC_advRobust_dev			
	Acc_0	Acc_1	Acc_2	Score
Block strategy	44.35	22.02	5.06	18.01
Block strategy+EFR	**47.01(+2.66)**	**24.27(+2.25)**	**7.73(+2.67)**	**20.55(+2.54)**
	GCRC_advRobust_test			
	Acc_0	Acc_1	Acc_2	Score
Block strategy	44.10	22.57	5.21	18.19
Block strategy+EFR	**46.59(+2.49)**	**28.57(+6)**	**5.78(+0.57)**	**20.78(+2.59)**

Table 2: GCRC_advRobust 数据集的实验结果

如Table 2所示，在数据集GCRC_advRobust 上使用所提出的方法相比基线分别在验证集和测试集上将Score 指标提升了2.54% 和2.59%，证明了本文提出模型能够稳定提升模型鲁棒性。

5 基于重叠片段生成的优化方法

虽然在3.2中对文档进行了合理的片段划分，但是仍可继续改进，因为无交集的片段划分可能破坏证据的完整性，导致指代不明确和关键语义缺失，如Figure 5为缺少主语的样例，在第（10）句话中“但它有两类”中的“它”与其指代“流行音乐”被分开为两个独立片段。

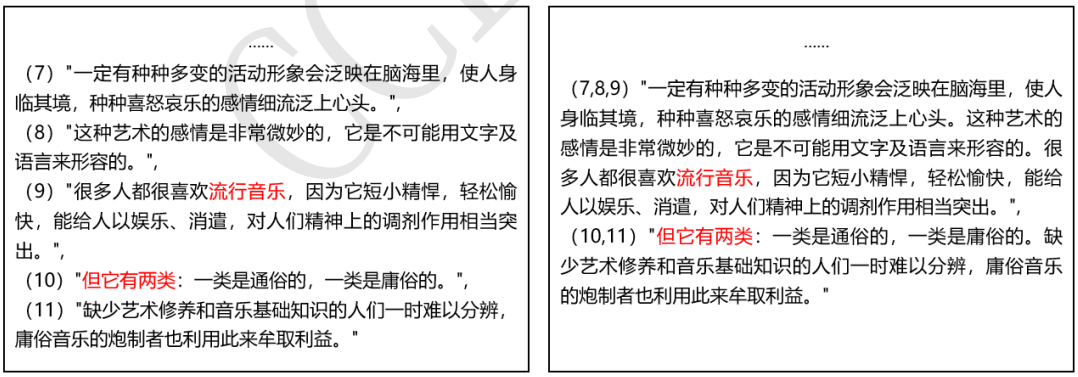
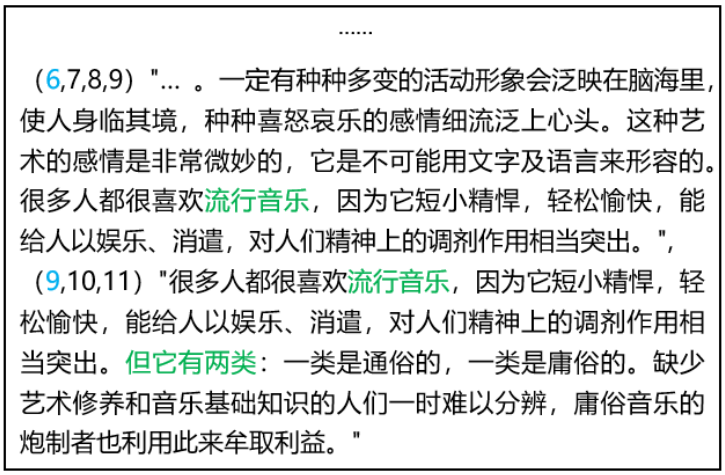


Figure 5: 原划分片段后的指代不明样例

为了生成更合理的文档片段，本文提出了重叠片段生成（**O**verlapping **S**egments **G**eneration）方法以优化原方法。在具体操作方面，该方法在原划分基础上，要求序号n $\geq$ 2 的文档片段包含其前一片段的后 γ 句，本文设置 $\gamma = 1$，在Figure 5的例子中，可将其划分为如Figure 6所示结果。



（b）有重叠方法划分后

Figure 6: 重叠片段生成划分后样例

其中括号中蓝色数字表示加入的有重叠句子，在标记为（9,10,11）的片段中绿色部分的“它”明确指代“流行音乐”。

5.1 实验结果

本文使用重叠片段生成方法（OSG）重新进行实验，其结果如Table 3所示。

模型方法	GCRC_advRobust_dev			
	Acc_0	Acc_1	Acc_2	Score
Block strategy	44.35	22.02	5.06	18.01
Block strategy+EFR	**47.01**	24.27	7.73	20.55
Block strategy+EFR（OSG）	46.73	**24.70(+0.43)**	**8.04(+0.31)**	**20.77(+0.22)**
	GCRC_advRobust_test			
	Acc_0	Acc_1	Acc_2	Score
Block strategy	44.10	22.57	5.21	18.19
Block strategy+EFR	46.59	28.57	5.78	20.78
Block strategy+EFR（OSG）	**48.26(+1.67)**	**31.60(+3.03)**	**6.25(+0.47)**	**22.26(+1.48)**

Table 3: 实验结果

在数据集GCRC_advRobust 上，如Table 3所示，使用重叠片段生成方法的自适应方法在所有指标下都得到了更好的效果，在两个数据集上将Score 指标在原片段生成方法基础上进一步提升了0.22% 和1.48%，这表示自适应方法确实能够提升模型的鲁棒性。

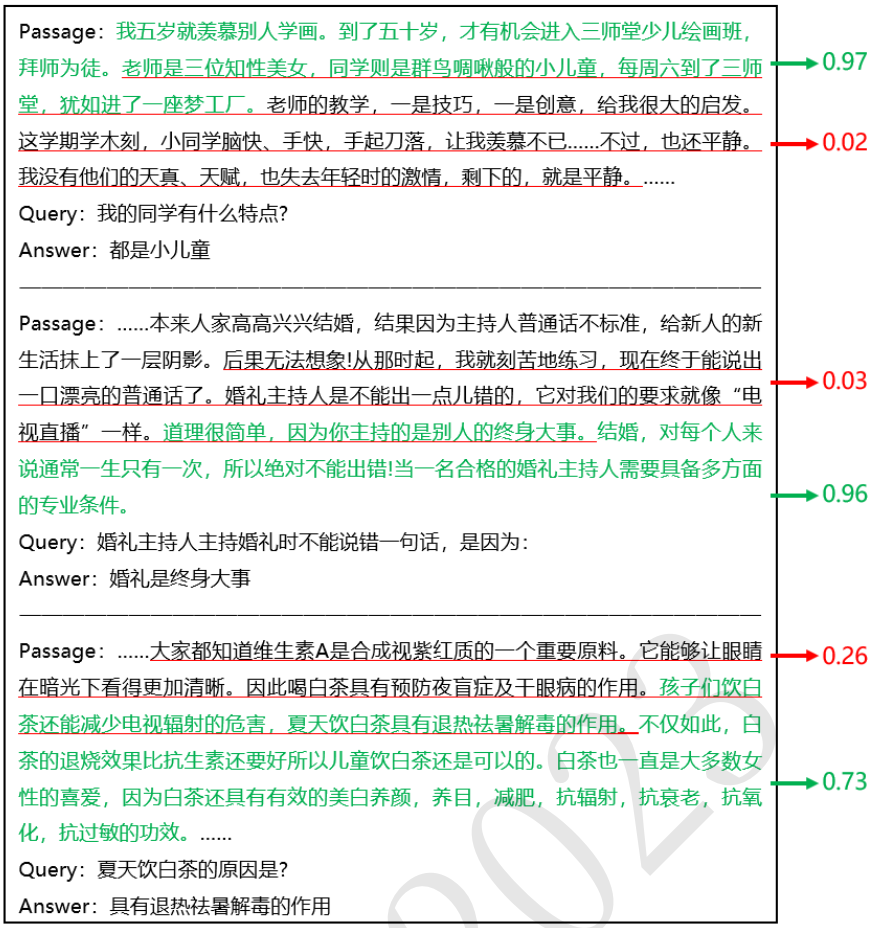


Figure 7: 自适应证据抽取效果样例

此外Figure 7也展示了端到端证据抽取方法的具体效果样例，重点在展示证据片段生成方法是有重叠片段划分并且证据句被包含在两个文章片段的结果，Figure 7中有红色下划线的语句表示与证据句有重叠的文章片段，绿色文字表示本文方法最终抽取的证据片段，右侧数据表示两部分结果证据片段抽取在3.2节中得到的置信分数，能够看出证据句被准确抽取出来，并且当证据被包含在两个文章片段中时会优先抽取出更具有完整语义的片段（即给片段较高置信分数）。

6 结果与讨论

证据抽取能提供原文中能够回答问题的证据，本章基于文章证据句子的精读能够优化模型性能的特点，提出了将关键（证据）重读模块引入文章整体训练中，自适应的改变证据片段的抽取，在无需增加参数的条件下提升模型性能上限。在具体实现上，该任务通过文章片段生成、证据片段识别和证据片段重读三部分完成关键重读模块的构建。

参考文献

Ahsaas Bajaj, Pavitra Dangati, Kalpesh Krishna, Pradhiksha Ashok Kumar, Rheeya Uppaal, Goldman Sachs, Bradford Windsor, Eliot Brenner, Dominic Dotterrer, Rajarshi Das, et al. 2021. Long document summarization in a low resource setting using pretrained language models. *ACL-IJCNLP 2021*, 120(4,268):71.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Minghan Li and Eric Gaussier. 2021. Keybld: Selecting key blocks with local pre-ranking for long

document information retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2207–2211.

Weikang Li, Wei Li, and Yunfang Wu. 2018. A unified model for document-based question answering based on human-like reading strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. Gcrc: A new challenging mrc dataset from gaokao chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330.

Jiawei Wang, Hai Zhao, Yinggong Zhao, and Libin Shen. 2021. What if sentence-hood is hard to define: A case study in chinese reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2348–2359.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Chen Zhang, Yuxuan Lai, Yansong Feng, and Dongyan Zhao. 2021. Extract, integrate, compete: Towards verification style reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2976–2986.