多模块联合的阅读理解候选句抽取*

吉宇[1,‡] 王笑月[1,‡] 李茹[1,2,*†]郭少茹[1,‡] 关勇[1,‡]
[1]山西大学　计算机与信息技术学院
[2]山西大学　计算智能与中文信息处理教育部重点实验室
[‡]{jiyu0515, wangxy0808,guoshaoru0928,guanyong0130}@163.com
[*]liru@sxu.edu.cn

摘要

机器阅读理解作为自然语言理解的关键任务，受到国内外学者广泛关注。针对多项选择型阅读理解中无线索标注且涉及多步推理致使候选句抽取困难的问题，本文提出一种基于多模块联合的候选句抽取模型。首先采用部分标注数据微调预训练模型；其次通过TF-IDF递归式抽取多跳推理问题中的候选句；最后结合无监督方式进一步筛选模型预测结果降低冗余性。本文在高考语文选择题及RACE数据集上进行验证，在候选句抽取中，本文方法相比于最优基线模型F1值提升3.44%，在下游答题任务中采用候选句作为模型输入较全文输入时准确率分别提高3.68%和3.6%，上述结果证实本文所提方法有效性。

关键词： 机器阅读理解 ； 候选句抽取 ； 递归抽取

Evidence sentence extraction for reading comprehension based on multi-module

Yu Ji[1,‡] Xiaoyue Wang[1,‡] Ru Li[1,2,*] Shaoru Guo[1,‡] Yong Guan [1,‡]
[1]School of Computer and Information Technology，Shanxi University
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education,Shanxi University,Taiyuan,China
[‡]{jiyu0515, wangxy0808,guoshaoru0928,guanyong0130}@163.com
[*]liru@sxu.edu.cn

Abstract

As a key task of natural language understanding, machine reading comprehension has been widely concerned by scholars at domestic and foreign. In order to solve the problem of multiple choice reading comprehension, which is difficult to extract evidence sentences due to the absence of clue annotation and questions involve multi-hop reasoning, we proposes a model of evidence sentence extraction based on multi-module combination. Firstly,we use some labeled data to fine-tune the pre-training model; secondly, the evidence sentences in the multi-hop reasoning problem are extracted recursively through TF-IDF; finally, the unsupervised method is combined to further filter the model prediction results to reduce redundancy. This paper is verified on the Chinese Gaokao and the RACE data set. In the extraction of evidence sentences, compared with the optimal baseline model, the F1 value of the method in this paper is increased by 3.44%. The accuracy of using evidence sentences as model input in

* 基金项目：国家重点研发计划重点专项(No.2018YFB1005103);国家自然科学基金(No.61772324)
† 通讯作者.

downstream answer tasks is 3.68% and 3.6% respectively higher than that of full text input. The above results confirm the effectiveness of the proposed method.

Keywords: Machine reading comprehension , Evidence sentence extraction , Recursive extraction

1 引言

随着自然语言处理技术的发展，国内外对于机器阅读理解的研究不断深入。本文重点关注机器阅读理解中的多项选择题任务(Mostafazadeh et al., 2016; Lai et al., 2017)，即：给定文章、问题和选项，要求根据文章回答问题，从多个选项中选择最佳选项。

对于该任务，研究者通常将整篇文章、问题及选项作为输入(Wang et al., 2018; Ran et al., 2019)并在三者之间两两交互，进行信息整合继而选出最佳选项。然与片段抽取式阅读理解不同，多项选择的答案难以直接从给定的文章中提取(Wang et al., 2019)，（如在RACE(Lai et al., 2017)中87%的问题不能直接从文章中找到答案），且并非全文都与当前选项有关。若将全文信息编码将引入噪声从而影响模型预测。特别对于文章较长的数据集（如高考语文阅读理解多项选择题，其平均长度为1,134.15字/篇），输入全文将导致模型很难提取出与问题及选项相关的“重要信息”，且答题缺乏可解释性。因此，为提取问题所需的“重要信息”并提升模型性能，本文针对多项选择阅读理解任务中的候选句抽取问题进行研究。

针对以上问题，Zhang et al. (2019)提出DCMN+，在对文章编码前加入候选句筛选工作，利用余弦相似度评估文章与选项间关联程度，以此缩短文章范围。但多项选择题的候选句通常为多句(Wang et al., 2019)，存在某些候选句与选项之间重叠度较低的情况，如图1所示，结合问题可得出S24包含判断选项A正误的关键信息，但从表面看S24与选项A关联度较低。若通过余弦相似度、模式匹配的方式查找，该类候选句很难抽出且会对后续答题造成影响。鉴于此，Trivedi et al. (2019)将候选句判断视为文本蕴含任务(Korman et al., 2018)，文章中句子视为前提、选项视为假设，判断两者之间是否存在蕴含关系。由于缺少标注数据，采用SNLI(Bowman et al., 2015)对模型进行微调之后进行预测。但该种方法未考虑候选句之间信息冗余的情况，如图1所示，通过语义相似度计算，可在文中抽取出与选项A的关联句S1与S15,但这两句所含的信息相同，对答案预测并无提升作用且增加了计算量。对此，Yadav et al. (2019)提出ROCC，从候选句集对选项及问题的信息覆盖度、候选句与选项及问题之间的语义相关性以及候选句之间冗余性三方面计算ROCC得分，进一步筛选抽取结果。在一定程度上缓解冗余现象，有效提高了候选句抽取的精确性。

上述方法的提出虽使模型性能获得巨大提升，但仍存在一些挑战。（1）直接将选项与问题拼接，忽视其拼接结果是否为一个完整的陈述句或存在语法错误，会对模型句意理解造成影响，如图1所示；（2）通常数据集中候选句在全文所占比例较小而无关信息占比较大，存在正负样本不均衡的情况；（3）对于需要多步推理的问题（即：A推B，B推C），判断选项正误与否的候选句与选项之间并不存在直接关联，需寻找选项候选句的候选句。

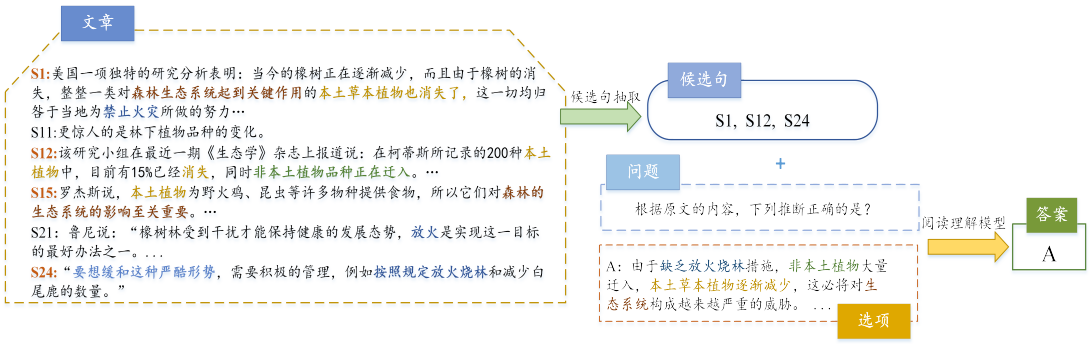

图 1. 高考阅读理解选择题候选句示例

面对上述挑战，本文在RACE和高考语文数据集（见5.2节）上进行实验。通过对数据集研究发现，文章候选句与选项之间的关联度较低。若仅用句对间关联较为明显的SNLI等数据集

训练或采用无监督方式，都无法较准确完整地将候选句抽出。故本文人工标注部分候选句对。考虑到选项信息不完整对候选句抽取的影响，本文将所有问题与选项拼接改写确保其不含语法错误；之后，使用构造的数据集对BERT(Devlin et al., 2018)进行微调；针对正负样本不均衡现象，采用FocalLoss(Lin et al., 2017)作为损失函数，在训练时推动模型更加关注于困难样本，降低简单负例的学习度，从而在整体上提高候选句抽取的F1值，基于此得到初步候选句集；对于多步推理问题导致候选句难以直接抽取的现象，本文提出基于TF-IDF的递归式抽取方法，进一步提升模型召回率；为保证候选句抽取结果的精确性，减少候选句之间的冗余，采用ROCC(Yadav et al., 2019)过滤重复信息，提升精确率。

为进一步评估候选句抽取质量及所提方法对后续答题的帮助，本文将抽取出的候选句集拼接，采用BERT与co-matching模型分别在RACE、高考语文阅读理解选择题数据集上进行实验，实验结果表明采用候选句集作为输入相比全文在高考及RACE数据集上分别提升了3.68%及3.6%。在候选句抽取上，本文所提方案相比于基线F1值进一步提升了3.44%及3.95%。

2 相关工作

候选句抽取工作，依据训练方式可划分为四种类型。（1）使用无监督方法为候选句抽取提供了指导，同时减省人工标注的消耗(Yadav et al., 2019)；（2）有监督方法通过标注数据训练模型，从而实现下游任务中自动抽取候选句的目的。Trivedi et al. (2019)使用文本蕴含语料(Bowman et al., 2015; Williams et al., 2018)为训练候选句抽取模型。对于不提供候选句标注的数据集，研究者从结构化知识库(Speer et al., 2016)中选取相关线索知识，训练模型(Hao et al., 2017; Lukovnikov et al., 2017)；（3）使用信息检索进行候选句抽取工作，通过强化学习(Geva and Berant, 2018)或pagerank(Surdeanu et al., 2008)学习如何在缺少明确训练数据的前提下进行候选句抽取。或是使用注意力机制在文本与选项及问题之间交互，使文章中与选项和问题相关部分的注意力权重更大(Ran et al., 2019; Tang et al., 2019)；（4）通过人工定义规则，抽取出含有噪声信息的候选句，使用弱监督方式训练模型(Min et al., 2018)。上述工作，各有其贡献之处与意义，推动了模型在相应下游任务上的性能表现。本文所提工作着重在对上述工作疏漏之处进行强化，综合使用有监督与无监督方式，使抽取结果可评价并且提高抽取结果的精确性也减省了数据标注工作量。同时，对上述模型中未能考虑到的选项信息缺失问题以及正负样本不均衡也进行了相应处理。此外，本文针对多步推理问题提出了一种多步信息抽取方式，进一步提升了模型抽取效果。并在下游任务中验证了模型的有效性。

3 候选句抽取模型

本文提出一种新的候选句抽取模型，模型整体架构如图2所示。其主要包含四部分：（1）选项改写模块：融合选项与问题所涵盖的信息，确保其结果无语法错误；（2）候选句抽取模块：从文章中初步筛选出与判断选项正误有关的句子集合；（3）TF-IDF递归抽取模块：在前一步的基础上，使用TF-IDF作为引导，抽取多步推理问题候选句，避免关键信息遗漏；（4）筛选模块：在所得句子集合上进一步筛选，提高候选句抽取精确率，降低信息冗余。

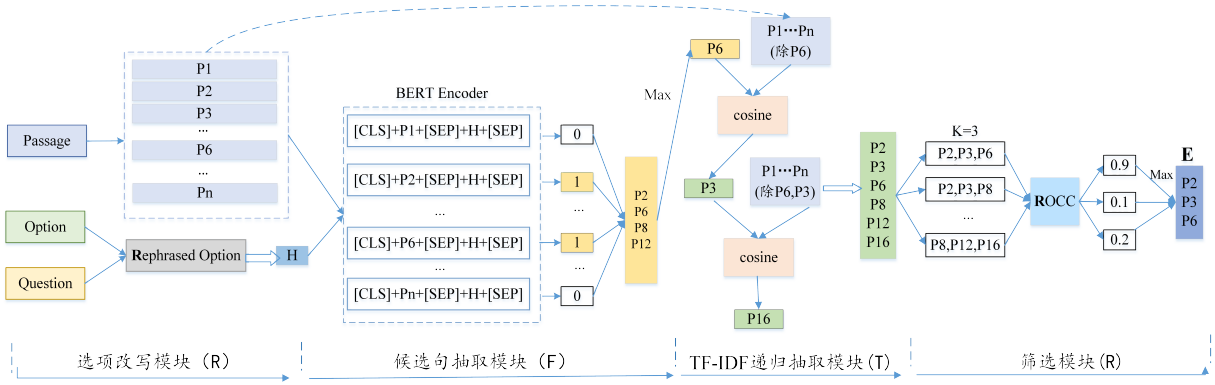

图 2. 候选句抽取模型

3.1 选项改写模块

通过对高考阅读理解及RACE数据集分析后发现，如图3所示，当问题为“下列说法符合（不符合）文意的一项是？（或其同义表述）,该类问题所蕴含的信息量较少，选项信息完整，无需对选项改写；而当问题为“下列对‘国外媒体关注点’的理解，不正确（正确）的一项是？”，选项内容为“科技竞争力”，若仅使用选项内容，其涵盖信息量过少，抽取对应候选句会较为困难；而若将问题与选项直接拼接，所得结果不符合语法规则。故需提取问题的关键信息，并将其与选项信息融合，形成一条完整的句子。

针对上述两种情况，首先采用正则表达式进行选项内容改写，使其形成完整陈述句$H = \{H_1, H_2, \ldots, H_m\}$其中$m$为选项改写句的长度；之后将文章切分为句子$P = \{P_1, P_2, \ldots, P_n\}$其中$P$为文章，$n$为文章中句子数量。

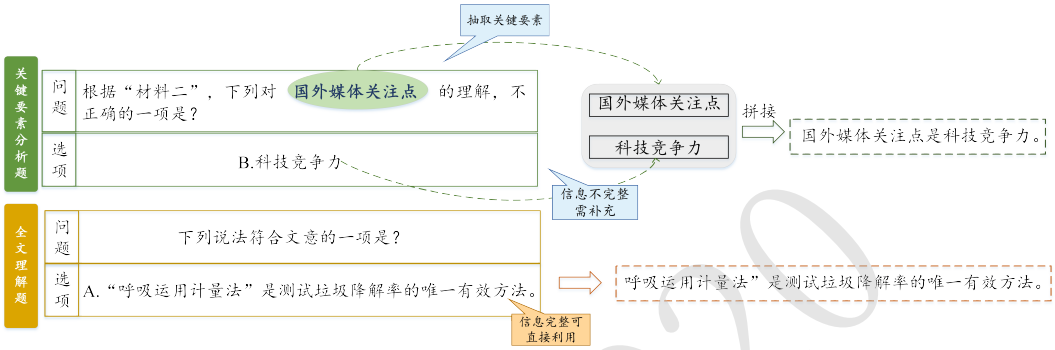

图 3. 选项改写示例

3.2 候选句抽取模块

该模块通过计算P_i与选项改写句H的关联度，初步抽取出候选句。本文在BERT基础上进行实验，首先将$[CLS]$,句子P_i,$[SEP]$,选项改写H,及$[SEP]$拼接后输入模型中，其中$[SEP]$为BERT中的片段分隔符，$[CLS]$为特殊字符（输入整体表示）。编码后，取$[CLS]$的编码结果$O_i \in R^d$进行分类，d为BERT隐藏层维度。

3.2.1 Focal Loss

由于候选句数据集中存在正负样本不均衡现象（RACE候选句数据集中，正负样本比为1：10）。本文采用FocalLoss作为损失函数，使模型聚焦于正样本的学习,缓解样本类别不均衡带来的风险。

输入的候选句对为(P_i, H),模型预测结果为P=[p_0,p_1],真值为Y=[y_0,y_1]。对于传统的交叉熵损失而言，其表示为：$CE = -(y_0 \log(p_0) + y_1 \log(p_1))$,显然，当负样本占比较大时，模型的训练会被负样本占据，使得模型难以从正样本中学习。

$$L_{fl} = \begin{cases} -\alpha(1-y')\log'_y & y = 1 \\ -(1-\alpha)y^{\gamma}\log_{1-y'} & y = 0 \end{cases} \tag{1}$$

FocalLoss在原有的基础加入权重系数γ及α,γ减少易分类样本的损失使模型更关注于困难的、错分的样本；α用于平衡正负样本本身数量比例不均,由此缓解了正负样本不均衡的现象。

3.3 TF-IDF递归抽取模块

由于阅读理解多项选择题中存在多步推理问题，如图1所示，该情况难以直接使用文本蕴含方式将选项对应候选句全部抽出。考虑到多步推理问题中存在链式关系，故基于上一步所得结果$E = \{E\}$，首先选出与选项改写句关联度最高的句子作为第一跳候选句$hop1$。继而，计算其与文章句子（除本身之外）的相似度，取相似度最高的作为第二跳候选句$hop2$。之后，计算文章句子中与$hop2$之间的关联度（$hop1$与$hop2$除外），并取关联度最高句子作为第三跳候选句，以此类推，重复K次（K值视具体情况设定）。将所得的句子与候选句集合E合并。

3.4 候选句筛选模块

为提升抽取结果精确性，降低无关及冗余信息比重。本文使用ROCC对结果进一步筛选。首先，使用上一步的抽取结果E并对其进行全组合$\binom{n}{m}$，生成候选句集合G，其中n为抽取结果的总共句子数，m为组合单位（可依据具体情况调节大小）。之后，对每组集合分别从(1)集合内部的信息冗余度(2)集合对选项的信息覆盖率(3)集合与选项之间信息相关性三个角度计算得分。

3.4.1 冗余度

通过计算给定集合中句对间信息重合度，来确保候选句的多样性和信息互补性。得分越低的句子集合，信息冗余度越低。

$$O(G) = \frac{\sum_{g_i \in G} \sum_{g_j \in G} \frac{|t|g_i| \cap t|g_j||}{\max(t|g_i|, t|g_j|)}}{\binom{|G|}{2}} \tag{2}$$

其中G表示给定句子集合，g_i与g_j分别表示集合中的某一条句子，$t(g_i)$表示g_i所包含的词集合（去重后），$|t|g_i| \cap t|g_j||$表示g_i与g_j的共有词数量。

3.4.2 覆盖率

该模块用于衡量给定集合G对选项改写句H的词汇覆盖率,由H与集合G之间的共有词的IDF值加权平均得到。Coverage值越大，意味该集合包含选项改写句的信息越多。

$$C_t(H) = \bigcup_{g_i \in G} t(H) \cap t(g_i) \tag{3}$$

$$C(H) = \frac{\sum_{t=1}^{|C_t(H)|} IDF\left[C_t(H)\left[t\right]\right]}{|t\left(H\right)|} \tag{4}$$

其中$C_t(H)$表示选项改写句与集合之间的共有词。

3.4.3 相关性

使用BM25(Robertson and Zaragoza, 2009)计算给定集合G与选项改写句H的相关度。计算公式如下：

$$R(H, G) = \sum_{i}^{n} w_i \cdot R(h_i, G) \tag{5}$$

从上述三个角度分别计算出给定集合得分后，综合得分计算集合的ROCC值。

$$S(G) = \frac{R}{\varepsilon + O(G)} \cdot R(\varepsilon + C(H)) \tag{6}$$

如式6中所示，R为集合与选项改写句的relevance得分，O为集合的overlap值，$C(H)$为集合对选项改写句的coverage值，为避免计算中出现分子或分母为0的情况，添加ε作为平滑项，实验中设ε值为1。之后，选ROCC得分最大集合作为最终的候选句集合E_2。

4 答题模型

得到候选句集合后，将其句子拼接为文章C。之后同问题Q，选项O_i一起作为答题模型的输入。

$$A_i = f(C, Q, O_i) \tag{7}$$

$$L(A_t|C,Q) = -\log \frac{\exp(W^T \cdot A_t)}{\sum_{j=1}^{m} \exp(W^T \cdot A_j)} \tag{8}$$

式7中，$f(.)$表示模型编码过程，所得$A_i \in R^d$为文章，问题，选项的最终表示，其中d为模型维度。式8中，$W \in R^{d\times 4}$ 为参数矩阵，A_t为问题的正确选项。

5 数据集

5.1 候选句数据集

高考候选句数据集：由于缺少中文阅读理解候选句语料，本文从数据集中随机抽取500道题,对每个选项人工标注其候选句。标注规则为：对应每个选项，文章中与判断其正误有关句子标注为1，反之，标注为0。为确保数据标注质量，本文采取交叉验证的标注方式：将数据二等分，由四个同学两两一组进行标注，各组内同学标注的数据相同。标注后两组同学交换进行两轮校验，针对标注结果中不一致数据，由仲裁者仲裁进行第三轮校验，剔除无法确定的数据，若无异议，经三轮验证后，将所得标注结果确定为最终候选句集合，包含45，311句对。其中训练集，验证集，测试集包含数据量分别为：36,254，4,528，4,529。

RACE候选句数据集：本文采用Wang et al. (2019)标注的500道RACE mid-challenge部分候选句对，共34，736句对，其中训练集，验证集，测试集分别为27,790，3,473，3,473 。由于初，高中试题难度有所区别，在验证候选句抽取对答题的影响时，本文仅使用RACE数据集中的初中部分进行测试。

5.2 阅读理解多项选择题数据集

本文采用RACE数据集中mid-challenge部分进行实验，共收集18，364道问题，按8：1：1方式将数据划分给训练、验证、和测试集；此外本文同时收集了2005-2019年高考语文阅读理解选择题共7886道，与RACE采用同样方式划分。

6 实验设计与结果分析

6.1 模型评价指标

候选句抽取评价指标： 实验采用F1值、P(精确率)、R(召回率)来评估候选句抽取效果，计算公式如下：

$$P = \frac{TP}{TP+FP} \times 100\% \quad P = \frac{TP}{TP+FN} \times 100\% \quad F1 = \frac{2\times P\times R}{P+R} \times 100\% \tag{9}$$

答题模型评价指标:对于答题部分，采用accuracy作为模型性能评价指标。

6.2 参数设置

针对不同数据集的实验参数设置如表1所示。

实验	DataSet	epoch	max_length	batch	learning rate	K	m
候选句抽取	高考	3	128	32	3e-5	3	4
	RACE	3	128	32	3e-5	3	2
答题模型	高考	6	450	40	1e-5	—	—
	RACE	3	320	32	5e-5	—	—

表 1. 模型参数设置

6.3 实验结果与分析

6.3.1 基线模型性能比较

表2中展示各模型在高考及RACE候选句数据集上的效果，对于高考候选句数据集，从表中可看出BERTwwm的P值，R值及F1值均高于BERT-base；ALBERT-base抽取候选句虽P值较高，但R值相比于BERTwwm低17.71个百分点。故针对高考数据集，本文以BERTwwm为基

础进行改进，结果表明结合RFTR(本文方法)后，模型效果在P值上提升5.41个百分点，R值提升1.99个百分点，F1值3.44个百分点。对于RACE数据集基线模型中BERT-wwm取得最优效果，故在此基础上结合RFTR后，效果提升了3.95个百分点。以上所述验证了所提方法的优越性。

	高考			RACE		
Baseline Model	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
ALBERT-base	81.50	46.13	58.91	76.34	59.08	66.61
BERT-base	73.59	61.84	67.21	77.30	61.10	68.25
BERT-wwm	73.78	63.84	68.45	78.03	60.78	68.33
BERT-wwm+OR	77.29	65.34	70.81	83.10	63.77	72.16
BERT-wwm+OR+FL	76.94	65.81	70.94	82.58	64.10	72.18
RFTR-ROCC	76.86	65.87	70.94	81.40	64.85	72.19
RFTR	79.19	65.83	71.89	84.36	63.23	72.28

表 2. 基于高考语文和RACE的候选句抽取结果

6.3.2 候选句抽取消融实验

为进一步研究所提方案对实验结果的影响，在高考及RACE候选句数据集上分别进行了消融实验。如表2所示,改写选项后（即表中OR）在两数据集上模型F1值相比基线分别提升2.36及3.83个百分点，并且P值和R值也均有提升，表明改写选项使信息更完整，语义更通顺，这对模型的语义学习有很大帮助；之后针对数据集中正负样本不均衡现象，使用Focal Loss进一步使模型效果在F1值上分别提升0.13和0.02个百分点，其中对于高考R值提升0.47个百分点，表明更换损失函数后，模型对正样本学习的偏向性增强；使用TF-IDF相似度计算抽取需多步推理问题的候选句使模型召回率分别提升了0.06和0.75，表明该方案可以有效缓解多步推理问题的信息损失；最后，使用ROCC从候选句集之间的冗余度，候选句集对选项信息覆盖率和候选句与选项相关性三方面考虑，进一步对结果进行筛选，在两数据集上P值分别提升2.33个百分点和2.96个百分点。

6.3.3 候选句抽取效果验证

本文在高考阅读理解选择题与RACE数据集上进行验证，将抽出的候选句拼接作为新文章输入模型，效果如表3所示，其中EV(RFTR)表示使用候选句作为文章的方法，该实验结果证明了候选句抽取的有效性。

	高考		**RACE**	
模型	**dev(%)**	**test(%)**	**dev(%)**	**test(%)**
BERT(base)	32.61	30.33	55.91	57.66
BERT+EV(RFTR)	36.29	31.34	59.51	59.87
co-matching	35.27	32.36	47.54	42.22
co-matching+EV(RFTR)	36.29	35.39	50.65	46.40
ALBERT	29.06	28.05	65.26	67.08
ALBERT+EV(RFTR)	32.11	29.57	68.92	69.30
DCMN	30.83	30.24	48.16	49.53
DCMN+EV(RFTR)	32.64	32.12	50.53	52.19

表 3. 高考语文与RACE数据集答题模型对比结果

6.3.4 K,m对候选句抽取实验结果的影响

为比较实验中TF-IDF递归抽取模块(见3.3节)中K值及候选句筛选模块（见3.4节）全组合中m值对候选句抽取的影响，本文进行了参数对比实验。实验结果如图4,5所示：

由图4可知，在高考数据及RACE数据集中，随着跳数的增加，候选句的召回率逐渐提高，当K为3时召回率与F1值达到最优，表明当跳数为3时可有效缓解多步推理问题的信息损

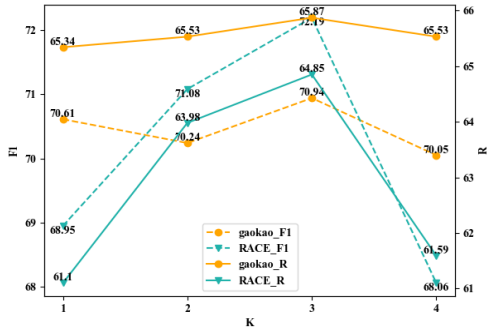

图 4. 递归抽取中K值变化

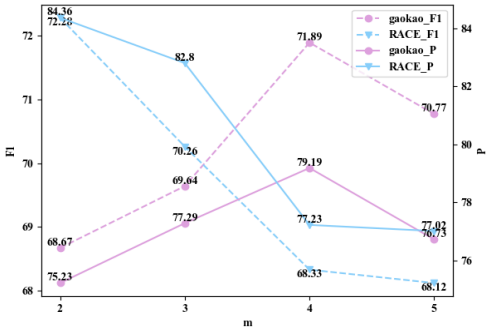

图 5. 筛选模块中m值变化

失；而当跳数为4时召回率下降，说明跳数过多也会引入一定的噪声。由图5可知，在高考数据及RACE中m值分别为4和2时精确率达到最优，说明ROCC可有效筛选冗余信息，最大限度地整合相关信息,从而剔除一部分无关句或冗余句。

6.3.5 错误抽取示例分析

本文选取了测试集中50条错误数据进行了分析，表4为列举的错误数据。

候选句	选项	预测结果	真实结果
蜉蝣这种生物大多数时间生活在水里，以藻类为食，当它们准备好繁殖，便爬出水面，在水边的植物上蜕皮，成为有翅的成虫。	蜉蝣有翅后即升空飞行。虽然飞行时间不长，但由此实现了生命的延续。	0	1
无论我们如何看待鲁迅，如何评价鲁迅先生的毕生之问和他为此所做的一切，现在，我们都依然得和他一起,承受一个各人心底诚信与爱都尚有不足的时代。	鲁迅的时代过去了，但那个时代的国民劣根性今天依然存在，为此我们要呼唤鲁迅，不要漠视鲁迅的存在。	0	1
从印刷的基本需求来看，排字机的字库通常要收7000多字。而从一般书报的需求来说，字体就有书版宋、报版宋、标题宋、仿宋、楷体、黑体...等十多种。	因字形字体的制约，汉字排版繁复。	0	1

表 4. 候选句抽取错误示例

由表可知，错误原因主要有以下三点：（1）指代问题，需辨别表中的“它们”指代“蜉蝣”，才可知“繁殖”与“生命延续”蕴含。（2）归纳概括问题：如“我们都依然得和他一起，承受一个各人心底的诚与爱都尚有不足的时代。”尚有不足的言外之意是：“但那个时代的国民劣根性今天依然存在”，然其表述差异性较大，导致计算机无法“理解”。（3）涉及归纳与知识融合：如需使模型知道“书版宋、报版宋、标题宋、仿宋等”即为“字形字体”。

7 结论与展望

本文针对多项选择阅读理解候选句抽取任务，以有监督方式抽取为基础，针对选项语义不完整、数据集正负样本不均衡、及抽取结果信息冗余等方面进行改进。在高考及RACE数据集上进行实验，证实了该方法的有效性。同时，还验证了候选句抽取对多项选择答案预测的帮助。此外，从表2中可看出，候选句抽取仍存在较大提升空间。结合错误分析，下一步计划挖掘阅读理解中更深层次的线索（如句间指代关联），提升候选句抽取效果，进一步提高答案预测的准确率。

参考文献

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Mor Geva and Jonathan Berant. 2018. Learning to search in long documents using document structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 161–176, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Yanchao Hao, Yuanzhe Zhang, Liu Kang, Shizhu He, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Daniel Z. Korman, Eric Mack, Jacob Jett, and Allen H. Renear. 2018. Defining textual entailment. *Journal of the Association for Information Science Technology*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations.

Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sren Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. In *International World Wide Web Conference 2017*.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option comparison network for multiple-choice reading comprehension.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations Trends® in Information Retrieval*, 3(4):333–389.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of ACL-08: HLT*, pages 719–727, Columbus, Ohio, June. Association for Computational Linguistics.

Min Tang, Jiaran Cai, and Hankz Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7088–7095, 07.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A co-matching model for multi-choice reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 746–751, Melbourne, Australia, July. Association for Computational Linguistics.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dan Roth. 2019. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dual co-matching network for multi-choice reading comprehension.