# 基于预训练语言模型先召回再验证的医疗知识阅读理解

陈旭

iamchenxu369@163.com

**摘要** 机器阅读理解通过让机器阅读和理解文本，来完成回答问题等相关任务，是自然语言处理的重要应用方向。2021年全国知识图谱与语义计算大会（CCKS2021）设置医疗科普知识阅读理解评测任务，要求评测系统针对用户所提出的搜索query，在相关的文章中找到对应的答案片段内容。本文提出一种基于预训练语言模型先召回答案片段、再验证答案句子的方法，通过增加对答案片段内句子句首的拟合，提升对长片段的召回效果，通过结合上下文的句子分类，提升答案预测的精度。在最终测试集上，本方法取得了答案抽取F1值0.6527的成绩，排名第一，验证了方法的有效性。

**关键词：**医疗知识阅读理解 预训练语言模型 长片段抽取 句子分类

## 1 引言

随着经济、社会发展，人们对健康问题越来越关注。互联网上有海量医疗资讯，帮助用户高效、准确地从中找到想要的信息，是技术研究人员面对的重要课题，对于提升人们的医学素养，也有积极作用。互联网信息以文本、图片、视频等多形式存在，自然语言处理就是研究文本相关问题的计算机科学分支，机器阅读理解是其中的一个重要应用方向，在搜索、问答、对话等系统中有广泛应用。随着深度学习和预训练技术的发展，机器阅读理解也取得了较大进步，但在实际应用中，仍有一些问题需要解决。

CCKS2021设置了面向中文医疗科普知识的内容理解评测任务，其中包含子任务医疗科普知识阅读理解。该任务针对用户所提出的搜索query，在相关的文章中找到对应的答案片段内容，以作为直接展示给用户的摘要，问题包括无答案问题和单个答案的问题，后者可能包含一个文本片段，也可能包含多个不连续的文本片段。除了上述特点和文本内容为医疗领域外，该任务答案片段相对较长，这些都是评测系统需要考虑和解决的问题。

本文针对医疗科普知识阅读理解子任务，提出了一种“先召回答案片段，再验证答案句子”的解决方法，在最终测试集上取得了答案抽取F1值0.6527的成绩，排名第一。本文方法的主要创新点在：

（1）在常用答案抽取模型的基础上，增加对答案片段内句子句首的拟合，有效保证了对长片段的抽取效果；

（2）在答案验证阶段，通过结合上下文对抽取片段中的句子做分类，进一步提高了答案的精度。

# 2 相关工作

## 2.1 机器阅读理解

机器阅读理解，根据任务形式的不同，主要分为四类：完形填空、片段抽取、多项选择和自由形式作答[1]。本文所述评测任务属于片段抽取阅读理解，该类任务公开的数据集也比较多，比如SQuAD2.0、CMRC2018等。在预训练语言模型出现之前，片段抽取类阅读理解的各种解决方法致力于细粒度的文本编码和让问题与原文更好地进行交互，然后利用交互信息预测答案起点和终点。预训练模型出现以后，将问题与原文拼接后通过预训练语言模型，就能很好的完成文本编码和信息交互融合。针对有不能回答的问题的情况，在前述答案抽取操作的基础上还要增加一个答案验证机制。比如根据抽取片段是答案的概率划分阈值，判断是否有答案；或者与片段抽取任务联合训练一个分类器；或者单独设计一个答案验证模块等[2]。本文所述评测任务还有一个特点是答案可能为多片段，目前公开数据集DROP[3]中包含多片段的答案抽取，其中典型的解决方法：一种是增加一个答案片段数目预测任务，然后从抽取的候选答案片段集合中，选取相应数目的不重叠片段作为最终答案[4]；另一种是将答案抽取建模为序列标注问题来解决[5]。

## 2.2 预训练语言模型

预训练语言模型是预训练方法在自然语言处理领域的应用，通过在大规模无标注数据上训练特定任务得到，谷歌提出的BERT[6]是其中典型代表。在许多自然语言处理任务上，使用预训练语言模型进行微调，只需要少量标注数据，就可以逼近、达到甚至超过传统深度学习的方法。此外，也有实验发现，如果使用通用的BERT模型效果不佳，那么在领域相关或任务相关的无标注数据上继续预训练，也能带来性能的提升[7]。MacBERT是BERT的一个改进版本，与BERT相比，主要改动是MLM任务使用相似词而不是[MASK] token来做mask。此外也使用了在其他BERT改进模型中证明有效的策略：全词mask、N gram mask、使用句子顺序预测任务而非下一句预测任务。在多个中文自然语言处理任务上，MacBERT表现出了优异的性能[8]。

# 3 方法

本文的方法包括答案召回和答案验证2个模块。针对每个问题，答案召回模块从原文中抽取20个可能性最高的答案片段，答案验证模块判断抽取片段中的每个句子是否属于最终答案。

### 3.1 答案召回

**模型结构**

在预训练语言模型出现以后，片段抽取类阅读理解的一种常用解决方法，是将问题与原文拼接，经过预训练语言模型后接2个全连接层，再做softmax操作，分别生成原文中某个位置是答案起点、终点的概率。针对有不能回答的问题的情况，再增加一个全连接层，利用经过预训练语言模型后[CLS] token对应的输出做二分类，判断该问题能否回答。当原文过长时，则将原文分成多段，分别与问题组合进行处理。

上述方法抽取的答案片段长度受到预训练语言模型最大句长的限制。本任务答案片段较长，经统计，训练集中有约5%的答案片段长度大于485，最大答案片段长度为3751。针对这个问题，另外考虑本任务答案片段多包含完整句子的特点，本方法的答案召回模块，对前述的常用方法作了如下改进：增加一个全连接层来拟合答案片段内句子句首位置（不包括整个答案片段的起始位置）。示例如表1，[Start]后面字符是答案起点，[End]后面字符是答案终点，[Mid]后面字符是答案片段内句子句首。一个答案片段会包括一个答案起点，一个答案终点，零个、一个或多个答案片段内句子句首。

**表 1.** 答案片段内句子句首示例

| 答案片段 |
| --- |
| 所以说，[Start]对于中暑的预防，最关键的还是减少外出活动，特别是在天气炎热的时候，即使是外出也要做好防护的工作，及时补充身体的水分和盐分。[Mid]夏天出门在外的时候，最好身边可以常备一些消暑的药物，比如藿香正气水，十滴水等常用的药品。[Mid]避免不时之需，同时避免在早上十点到下午四点之间外出，因为这段时间是最热的。[End]对于已经出现中暑的病人，我们应该对其做适当的治疗。 |

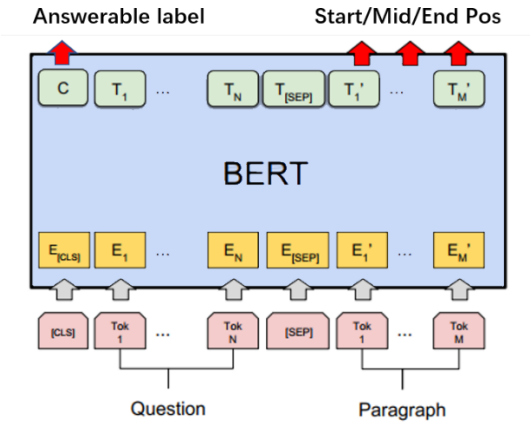


**图 1.** 答案召回模型

针对样本生成的某个输入模型的feature包含完整答案的部分片段但不包含答案起点或终点的情况，在模型训练阶段可以优化对答案片段内句子句首位置的拟合效果，在模型预测阶段可以通过答案片段内句子句首位置召回答案片段。图1是答案召回模型的示意图。

**优化目标**

模型的优化目标是最大化正确的答案片段起点、答案片段内句子句首、答案片终点位置的对数似然和“是否有答案”分类任务正确类别的对数似然。以答案片段起点为例，目标函数表示为：

$$\log\left(\frac{\sum_{k\in A} e^{s_k}}{\sum_{i=1}^{n} e^{s_i}}\right)$$

$A$是答案片段起点的集合，$n$是原文token数目，$s_i$是原文第i个token在做起点拟合的全连接层对应位置输出的logit。

**结果预测**

预测阶段，模型输出feature中原文每个位置起点logit、句首logit、终点logit。起点logit由大到小排序，取最大的n_best_size个logit对应的位置索引，得到起点索引列表，相同方法得到终点索引列表，取原文所有句首的索引，得到句首索引列表。

起点和终点索引列表、起点和句首索引列表、句首和句首索引列表、句首和终点索引列表中的元素，两两组合，排除不合理的组合后，得到候选答案片段起点、终点。根据候选答案片段起点、终点来源于哪个列表，定义起点类型、终点类型，并将起点、终点位置相应类型logit相加，作为这个片段的得分。按得分高低对片段排序后，保留前n_best_size个片段结果。对保留片段的得分做softmax操作，得到每个片段是答案的概率值。

把n_best_size个结果中，两片段相邻且前面片段的终点类型为“句首”、后面片段的起点类型为“句首”的片段合并，合并后片段是答案的概率值取原来两片段概率值中的大值。根据片段去除标点符号后的内容，对片段去重。取去重后概率值由大到小前20的片段，作为召回模型的最终输出结果。

### 3.2 答案验证

**模型结构**

验证模型对召回模型召回的答案片段中每个句子做分类，预测其是否属于答案。召回模型处理了答案片段过长的问题，同时对于答案片段的起点、终点在句子内部的情况也能覆盖到。而验证模型则可以处理问题无答案和有多个答案片段的情况。

通过观察部分数据，发现问题主要是关于某个疾病的症状、病因、治疗等方面，答案多是较长的一段或几段文本。单独看答案中某个句子，有的与问题关联明显，有的则不太明显，因此在做这个分类任务时，没有将答案句子单独

与问题做匹配，而是将抽出的整个答案片段与问题做交互融合，再对其中每个句子做分类，这样做，使得句子分类可以利用到上下文信息。

验证模型的结构是在预训练语言模型后接2个全连接层并做softmax操作，一个全连接层用来做句子级的分类，判断句子是否属于答案；一个用来做token级的分类，判断token是否在答案中。模型的输入是将问题与答案片段拼接，这个答案片段是由原始抽取出的答案片段分句后再用[SEP] token连接得到。句子分类使用的是该句子前面的[SEP] token对应的预训练语言模型最后一层的输出，token分类使用该token对应的预训练语言模型最后一层的输出。图2是答案片段只有2个句子时的验证模型示意图。

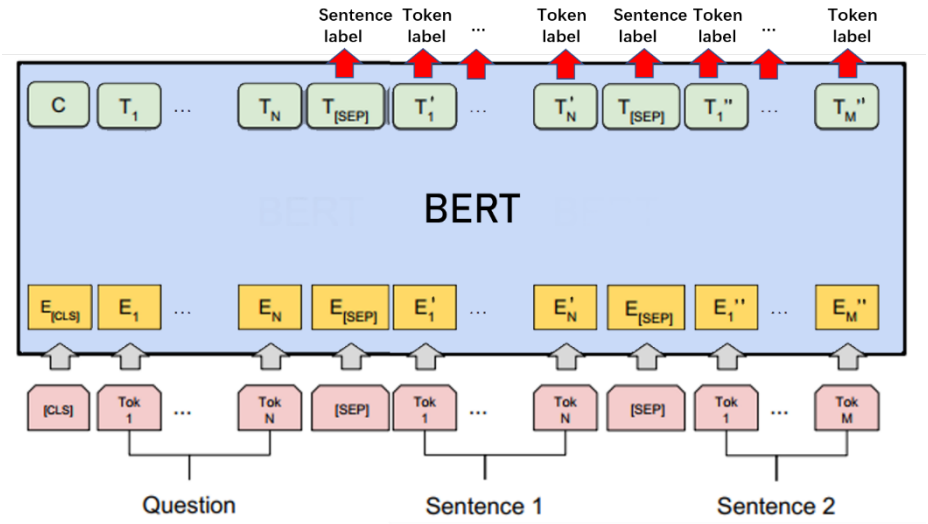


**图 2.** 验证模型结构

**优化目标**

模型的优化目标为最小化句子分类任务的交叉熵与token分类任务的交叉熵之和。

**结果预测**

预测阶段，利用句子级分类层得到一个句子属于答案的概率sent_ans_prob；将这个句子所有token在token级分类层输出的logit作平均，平均值再做softmax操作，得到融合token分类结果给出的句子属于答案的概率token_ans_prob。由于数据处理阶段对一个问题的所有片段提取结果是按照4个句子（4gram）为单位去重，所以里面会有单个句子重复出现的情况，这里将每个句子的上述2个概率，更新为与其重复的所有句子概率的平均值。

使用单折模型预测时，将sent_ans_prob与token_ans_prob加权求和，再根据阈值，确定一个句子是否为答案；使用多折模型预测时，先将多折的

sent_ans_prob和token_ans_prob分别求平均，平均值作为最终sent_ans_prob和token_ans_prob，然后二者加权求和并与阈值比较，得到预测结果。对一个问题预测出的全部答案句子按单句去重后得到最终答案。比赛中最佳权重和阈值是通过下述方法确定的：在0到1的范围内，以0.01为步长分别得到权重和阈值的候选取值，遍历所有的权重和阈值组合，在线下测试集上句子分类的F1值最高时的组合作为最佳值组合。

# 4 实验

## 4.1 答案召回

**数据处理**

原始训练集中，个别数据存在答案片段重复、重叠的情况，少量数据存在2个答案片段在原文中是连续的情况，对这些情况统一处理，都合并成一个答案片段，示例如表2。

**表 2.** 原始数据处理示例

| | 答案片段 |
|---|---|
| 合并处理前 | {<br>…<br>"answers": [[<br>"从而引发肝部囊肿出现，严重者可伴有多囊肾产生，加重患者的病情。",<br>"小部分患者因为创伤、炎症以及结石等因素，严重堵塞到了小胆管，继而形成了囊肿，伤害到患者的肝区健康。“<br>]]<br>} |
| 合并处理后 | {<br>…<br>"answers": [[<br>"从而引发肝部囊肿出现，严重者可伴有多囊肾产生，加重患者的病情。小部分患者因为创伤、炎症以及结石等因素，严重堵塞到了小胆管，继而形成了囊肿，伤害到患者的肝区健康。"<br>]]<br>} |

原始训练集处理后，取1/10作为线下测试集，剩余部分按照10折交叉验证划分训练集和验证集。

**实验设置**

预训练语言模型使用开源的MacBERT-large再训练模型[1]；最大句长为512；原文切分为多个片断时，相邻片段重叠长度为128；训练的批大小为4，训练4步做一次梯度更新；训练5轮；优化器使用AdamW，权重衰减设置为0.01；最大学习率设置为3e-5，使用学习率预热和线性衰减的策略，预热比例设置为0.1；预测阶段n_best_size参数的设置与原文长度有关，原文长度大于等于2000时为200，原文长度大于等于1000、小于2000时为150，原文长度小于1000时为100。

**实验结果**

本文从2个方面评估答案召回的效果：一是对一个有答案的问题，从其对应的模型抽取的所有答案片段中，选择句子进行组合，组合结果与该问题的正确答案计算F1值，得到最高F1值，然后求所有有答案问题最高F1值的平均值（最高F1平均值）；二是对一个有答案的问题，模型抽取的所有答案片段去重句子数与该问题对应的原文的去重句子数的比值，求所有有答案问题的这个比值的平均值（句子数比值平均值）。在线下测试集上，单折模型和10折模型融合的答案召回效果如表3。

**表 3.** 答案召回模型效果评估

| 方法 | 最高F1平均值 | 句子数比值平均值 |
|---|---|---|
| 单折模型 | 0.9760 | 0.7267 |
| 10折模型融合 | 0.9791 | 0.7226 |

## 4.2 答案验证

**数据处理**

答案验证模块数据处理主要包括3项内容：一、给句子打标签，标记是否属于答案，做法是对一个问题抽取的答案片段中的每个句子，将其与正确答案中的每个句子分别计算F1值，如果最大的F1值大于0.85，则将这个抽取句子标记为正例，否则为负例；二、对一个问题的全部抽取结果，按照4个句子（4gram）为单位去重，如果两个4gram片段F1值大于0.85，则认为是重复的，保留第一次出现的4gram；三、对于无答案问题的抽取结果，只保留在有答案问题的抽取结果中出现的句子用来训练模型。

答案召回模块10折交叉验证中所有验证集的抽取结果合并，再按照10折交叉验证拆分，经过上述3步处理后，作为本模块训练集和验证集。答案召回模块的线下测试集通过10折模型融合后的抽取结果，经过上述前2步处理，作为本模块的线下测试集。

[1] https://huggingface.co/luhua/chinese_pretrain_mrc_macbert_large

**实验设置**

使用与答案召回模块相同的预训练语言模型；最大句长为512；训练的批大小为4，训练4步做一次梯度更新；训练5轮；优化器使用AdamW，权重衰减设置为0.01；最大学习率设置为3e-5，使用学习率预热和线性衰减的策略，预热比例设置为0.1。

**实验结果**

本文在线下测试集上，分别评估了5种处理方法下句子分类的F1值和预测答案与参考答案的F1值，这5种方法分别是：使用单折sent_ans_prob做分类、10折融合后使用sent_ans_prob做分类、10折融合后使用sent_ans_prob与token_ans_prob的加权做分类、10折融合后使用sent_ans_prob与token_ans_prob的加权（权重、阈值来自线下测试集经过数据处理第3步后句子分类最高F1对应取值）做分类、第4种处理方法在加权前概率更新为同问题答案中与其重复的所有句子概率的平均值。结果见表4。

表 **4.** 答案验证模型效果评估

| 方法 | 句子分类F1 | 答案抽取F1 |
|---|---|---|
| 单折sent_ans_prob | 0.6984 | 0.6211 |
| 10折融合sent_ans_prob | 0.7158 | 0.6362 |
| 10折融合sent_ans_prob+token_ans_prob | 0.7158 | 0.6364 |
| 10折融合sent_ans_prob+token_ans_prob+drop_part_na | 0.7154 | 0.6382 |
| 10折融合sent_ans_prob(avg)+token_ans_prob(avg)+ drop_part_na | 0.7195 | 0.6366 |

采用上述第5种处理方法，本文方法在比赛最终测试集上，答案抽取F1值为0.6527。

## 5 总结

本文针对CCKS2021医疗科普知识阅读理解子任务，提出了一种“先召回答案片段，再验证答案句子”的解决方法。答案召回模块，在常用片段抽取模型的基础上，增加了对答案片段内句子句首的拟合，有效保证了对长片段答案的召回效果。答案验证模块，将召回片段中的句子放在上下文中与问题交互融合再做分类，从结果来看效果良好。分析本方法在线下测试集上句子分类的结果，发现对于句子中含有关键词，但句子内容不能回答问题的情况识别地不好，如果继续优化此方法的效果，需要针对这类问题想办法，如针对性数据增强、将预训练语言模型在医疗数据上做领域预训练、引入医疗知识等。另外，实验中，sent_ans_prob与token_ans_prob加权对系统效果基本没有提升，如何更好利用token级的信息，也值得做更多尝试。相信随着新方法的出现，本任务效果会不断提升。

# 参考文献


1. Chen, D.: Neural Reading Comprehension and Beyond. Ph.D. thesis, Stanford University (2018).
2. Zhang, Z., Zhao, H., and Wang, R.: Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond. arXiv preprint arXiv:2005.06249 (2020).
3. Dua, D., et al.: DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In: Proceedings of NAACL (2019).
4. Hu, M., et al.: A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning. In: Proceedings of EMNLP (2019).
5. Segal, E., et al.: A Simple and Effective Model for Answering Multi-span Questions. In: Proceedings of EMNLP (2020).
6. Devlin, J., et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL (2019).
7. Gururangan, S., et al.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of ACL (2020).
8. Cui, Y., et al.: Revisiting Pre-Trained Models for Chinese Natural Language Processing. In: Findings of EMNLP (2020).