清华与微软团队联合提出基于领域知识迁移学习的神经信息检索

原创 清华&微软团队 AI科技评论 2020-09-01

作者 | 清华&微软团队

编辑 | 陈大鑫

随着深度学习的快速发展，神经网络模型在CV、NLP等很多领域已经取得了显著超越传统模型的效果。然而，在信息检索领域，神经网络模型的有效性却仍然受到质疑。

例如这两年来，如BERT一样的预训练语言模型在很多自然语言处理的任务上取得了不错的效果，也成为了众多NLP任务的基线模型。然而，在信息检索领域，预训练语言模型在信息检索数据上的表现却并不突出。

那么，神经网络模型在信息检索领域的作用是否被夸大，又该如何有效地将神经网络模型应用于开放域信息检索场景？

基于上述问题，清华大学刘知远联合微软团队不仅提出了基于强化学习的弱监督数据筛选模型（ReInfoSelect）等相应解决方案，而且为了更好地解决开放域信息检索问题，刘知远老师所在的清华大学自然语言处理与社会人文计算实验室于近日开源了开放域信息检索工具包——

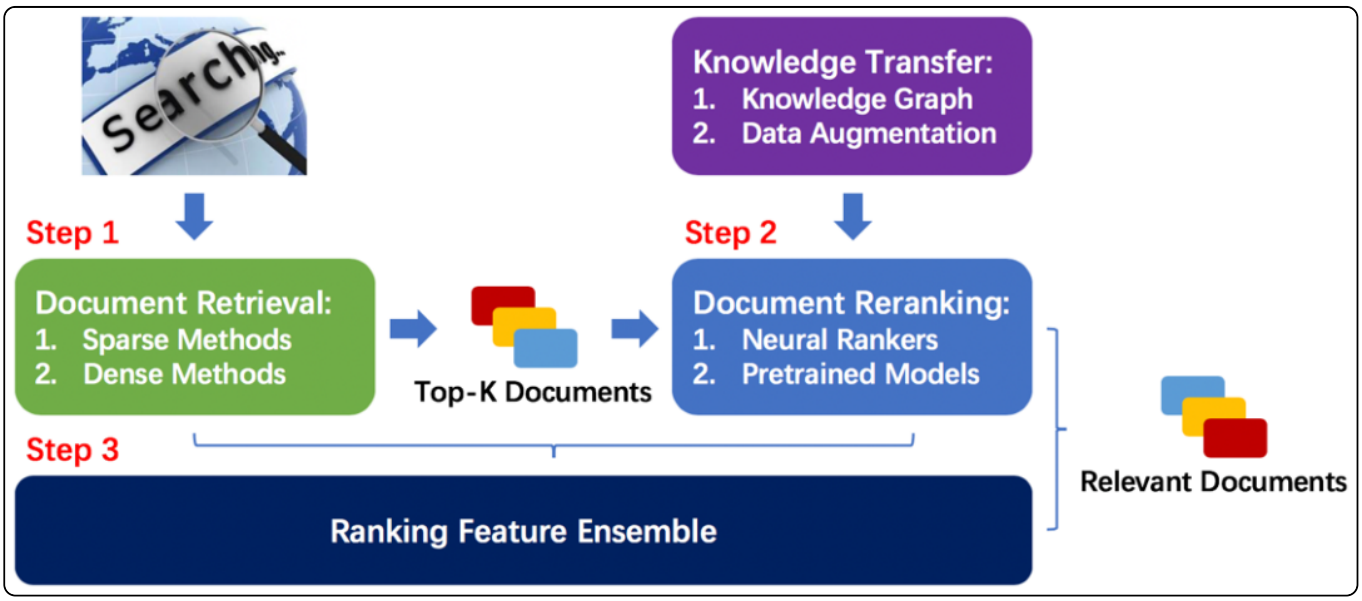
OpenMatch（基于Python和PyTorch开发），以及神经网络信息检索必读论文集——NeuIRPapers。

接下来，AI科技评论将和大家一起共探解决之道！

1

信息检索

信息检索顾名思义，就是根据用户给定的问题，通过模型的搜索返回相关的文档（如下图所示）。整个检索过程大致分为两步：文档检索、文档重排序。



信息检索框架

在文档检索中，由于检索规模巨大，且需要保证检索的召回率，因此，我们往往采用两类方法来实现。第一种为Sparse Methods，例如BM25,SDM等基于精确词语匹配的模型；第二种是Dense Methods，例如ANN(Approximate Nearest Neighbor)，其基于神经网络得到问题和文档的表示，通过计算二者表示的相似程度来对候选文档进行排序。

对于文档重排序而言，由于只需要对文档检索出来的前K个文档重排序，因此，我们更侧重于检索的精度和效果。

此类模型主要分为两类：

1、基于神经网络的信息检索模型，此种模型主要侧重于通过端到端训练，学习词向量之间的相关性特征，例如K-NRM，Conv-KNRM,TK等。

2、诸如BERT之类的预训练模型，此种模型通过预训练语言模型来增强模型的效果。

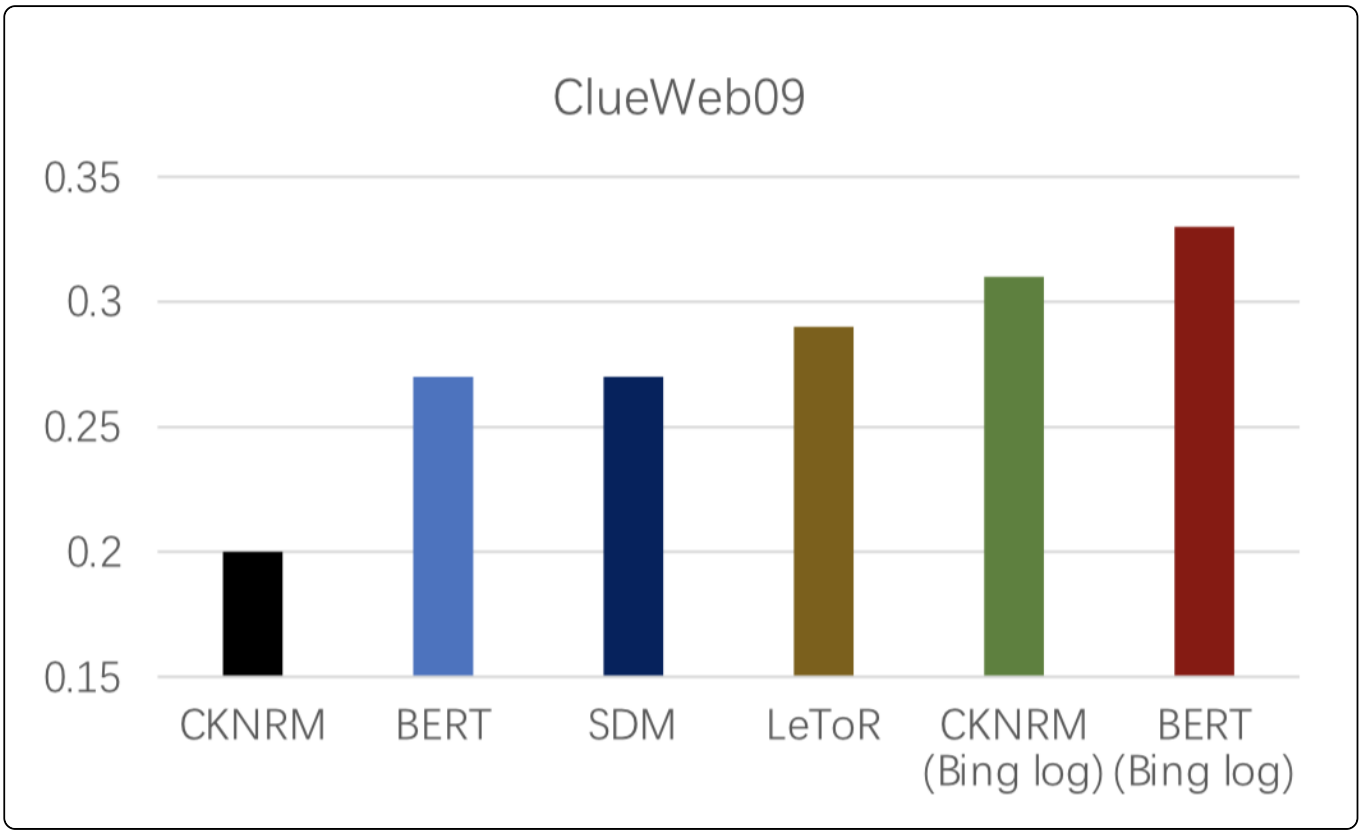
最后，我们可以通过Coordinate Ascent或RankSVM来整合两个阶段的全部模型特征，从而提升整体检索的效果。

在本文中，我们重点介绍第二步，以及如何更好的将领域知识迁移到文档重排序模型中。

— 2 —

预训练语言模型与信息检索

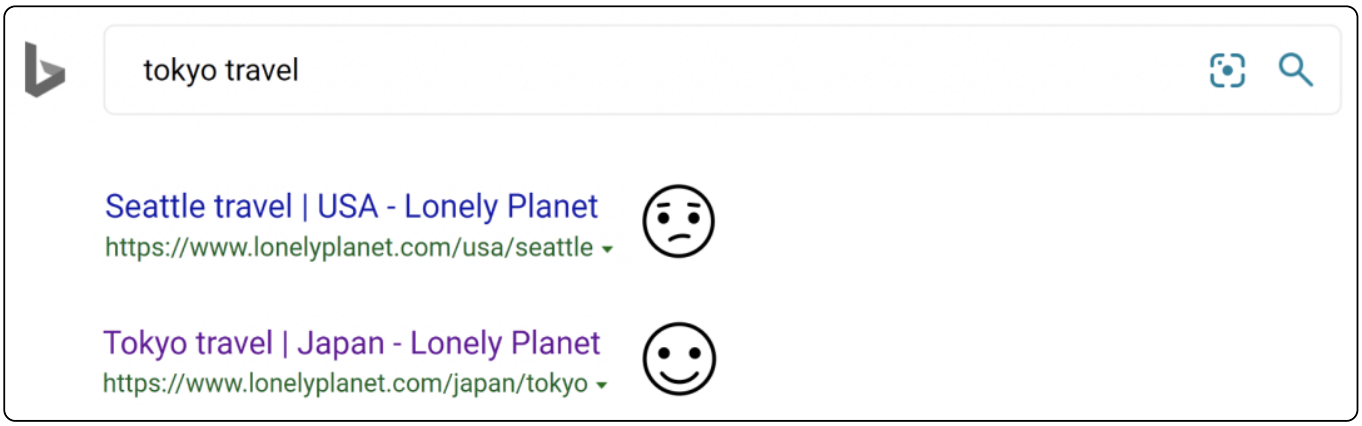
这两年来，如BERT一样的预训练语言模型在很多自然语言处理的任务上取得了不错的效果，也成为了众多NLP任务的基线模型。然而，在信息检索领域，如下图所示以ClueWeb09为例，预训练语言模型在信息检索数据上的表现却并不突出，没有显著优于基于传统信息检索特征的Learning to rank (LeToR)模型。但是通过Bing搜索日志训练的模型却得到了显著的提升，从而说明文本相关性监督信号对于神经网络信息检索是很重要的。



不同检索模型在ClueWeb09效果

因此，我们不禁会问为什么预训练模型对于我们的信息检索模型没有起到像其他任务一样的效果。我们来看下面一个例子。如下图所示，当用户想去搜索“Tokyo travel”的时候，是希望获取与东京

旅行相关的信息，而不希望得到其他地点的旅行信息。



信息检索样例

然而，考虑基于语言模型训练的预训练模型。如下图所示，以Masked Language Model为例，在这个例子中，“Seattle”和“Tokyo” 同属地名，均符合语言学规律，因此两个词都可以填入我们预先设置好的“[MASK]”位置，从而导致二者在词义表示上非常接近。

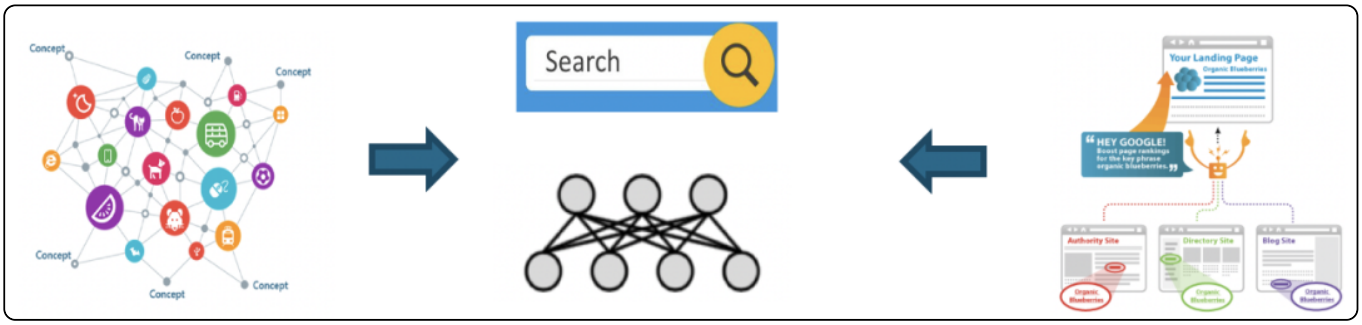
Masked Language Model样例

但是在信息检索场景中，我们并不希望“Seattle”和“Tokyo”这两个词的表示非常接近（毕竟西雅图和东京还是“有点”距离的）。因此，我们往往需要大量的文本相关性标签来端到端地训练我们的模型，从而学习到词语之间的匹配特征，以此提升神经信息检索模型的效果。

然而，文本相关性标签通常难以获得。一些相关性数据，例如搜索引擎用户的搜索日志，往往会涉及到隐私保护和商业机密，这对于研究人员来讲是难以获得的。

而如果通过招募人员进行大规模文本相关性标注，也需要大量的时间成本和人力成本，且对于一些特殊检索领域的标注，如医学、小语种等，对标注者的要求非常苛刻。以上的种种原因导致了文本相关性标签较少，从而使得神经网络模型在检索领域难以得到其应有的效果。

这时就需要领域知识迁移学习（如下图）派上用场，有两个方面可以实现它：融合外部知识、基于弱监督信号训练；这二者可以增强神经网络模型在信息检索模型中的效果。



基于领域知识迁移的神经信息检索模型

3

融合外部知识的神经信息检索模型

在信息检索场景中，用户检索词通常包含知识图谱实体信息，通过引入领域内知识图谱可以帮助我们提升检索模型的效果，帮助模型更好的理解相应实体的语义信息。

例如：给定“Obama family tree”这样一个问题，候选文档中往往出现一些相关实体，比如“United States”、“Michelle Obama”等等。通过融合知识图谱的相应的语义信息可以更好的增强实体的表示，从而提升信息检索的效果。

Query: Obama family tree

Documents:

Ranking	Document	Score
1	The family of **Barack Obama**, the 44th President of the **United States**, and his wife **Michelle Obama**...	12.0
2	**Barack Hussein Obama** is an American politician who served as the 44th ...	5.0

信息检索中的实体信息

以EDRM模型为例，其利用了知识图谱的三种语义信息来增强实体的表示，分别是实体嵌入式表示、实体描述信息、实体类型信息。

实体嵌入式表示通过知识图谱结构信息描述实体间的相似度。

实体描述信息往往包含实体最重要的语义，比如针对于实体“Obama”来说实体描述信息会给出“贝拉克·侯赛因·奥巴马，1961年8月4日生，美国民主党籍政治家，第44任美国总统，美国历史上第一位非裔美国人总统。”这样的解释，其往往反映了实体最重要的特性。

实体类型信息往往提供了一个机会来建立实体之间的隐含联系，比如：“Barack Obama”属于“Person”和“Leader”等类型。

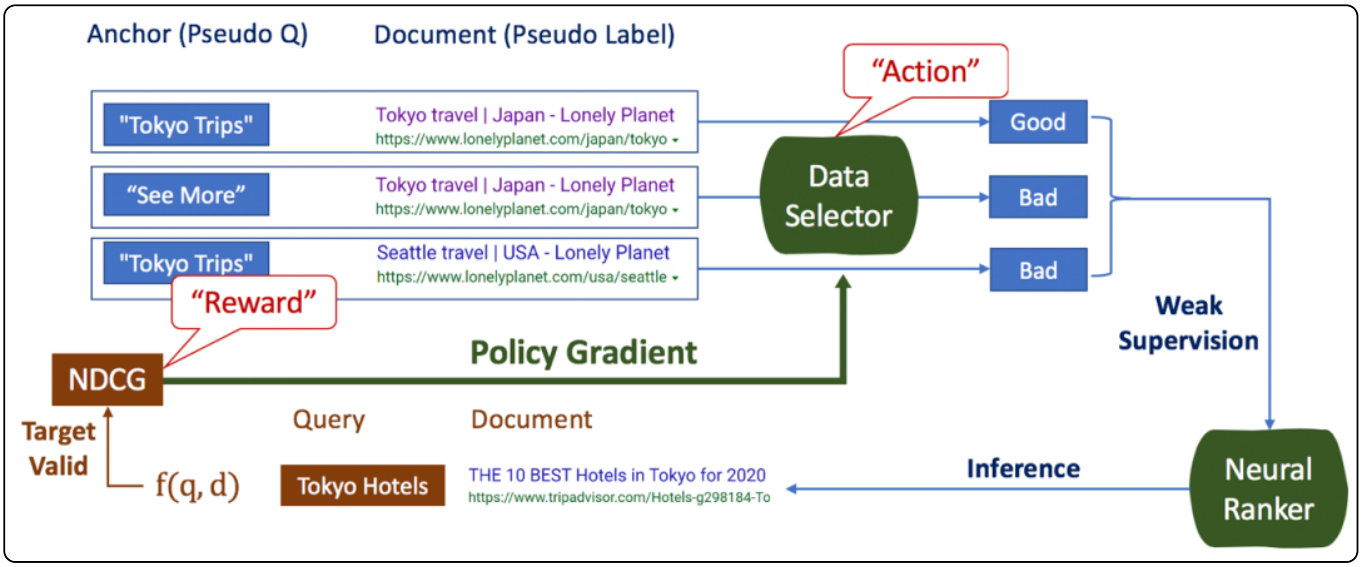
4

基于弱监督数据训练的神经信息检索模型

针对文本相关性标签少的问题，我们想到互联网中存在大量弱监督文本相关性标签，如：新闻标题-内容、网页锚文本-目标网页文本等。我们可以利用这些数据，结合负例采样，生成训练数据。

但这些弱监督数据通常含有大量噪声数据，例如：新闻标题通常为吸引读者眼球，与内容相关性并不高，俗称“标题党”。网页锚文本许多也并无实际意义——“see more”等。如果将所有数据一并训练，必然会对模型产生极大误导。

因此，我们提出了基于强化学习的弱监督数据筛选模型：ReInfoSelect。通过引入一个数据选择器来过滤噪声数据，训练神经网络信息检索模型，并通过检索模型在少量相关性标注数据上的表现来指导数据选择器的数据选择，从而提升数据筛选及文档排序效果。



ReInfoSelect框架

除了使用互联网中已有的相关文本作为弱监督文本相关性标签之外，我们还利用了训好的生成模型，基于现成的问答数据生成伪语料，构建弱监督文本相关性数据，训练神经网络信息检索模型。

5

开源工具及实验结果

为了更好地解决开放域信息检索问题，清华大学自然语言处理与社会人文计算实验室近日开源了开放域信息检索工具包——OpenMatch，以及神经网络信息检索必读论文集——NeuIRPapers。

OpenMatch是清华大学计算机系与微软研究院团队联合完成的成果，基于Python和PyTorch开发，它具有两大亮点：

1、为用户提供了开放域下信息检索的完整解决方案，并通过模块化处理，方便用户定制自己的检索系统。

2、支持领域知识的迁移学习，显著提升模型效果。

工具包地址：

https://github.com/thunlp/OpenMatch

论文集地址：

https://github.com/thunlp/NeuIRPapers

OpenMatch中提供的各模型如下表所示：

相关文档检索	文档重排序	领域知识迁移学习
BM25	K-NRM[1]	EDRM[3] (知识增强)
ANN	Conv-KNRM[2]	ReInfoSelect[8] (数据增强)
	TK[5]	
	Pretrained Models	
	Coor-Ascent	

表1 OpenMatch模型表

其中，Pretrained Models基于huggingface's Transformers实现。

地址：https://github.com/huggingface/transformers

OpenMatch采用Robust04，ClueWeb09-B和ClueWeb12-B13作为基准数据。对各模型效果进行测试(仅对数据集进行5折交叉验证，长文档截断取第一段文本)。

结果如下图所示：

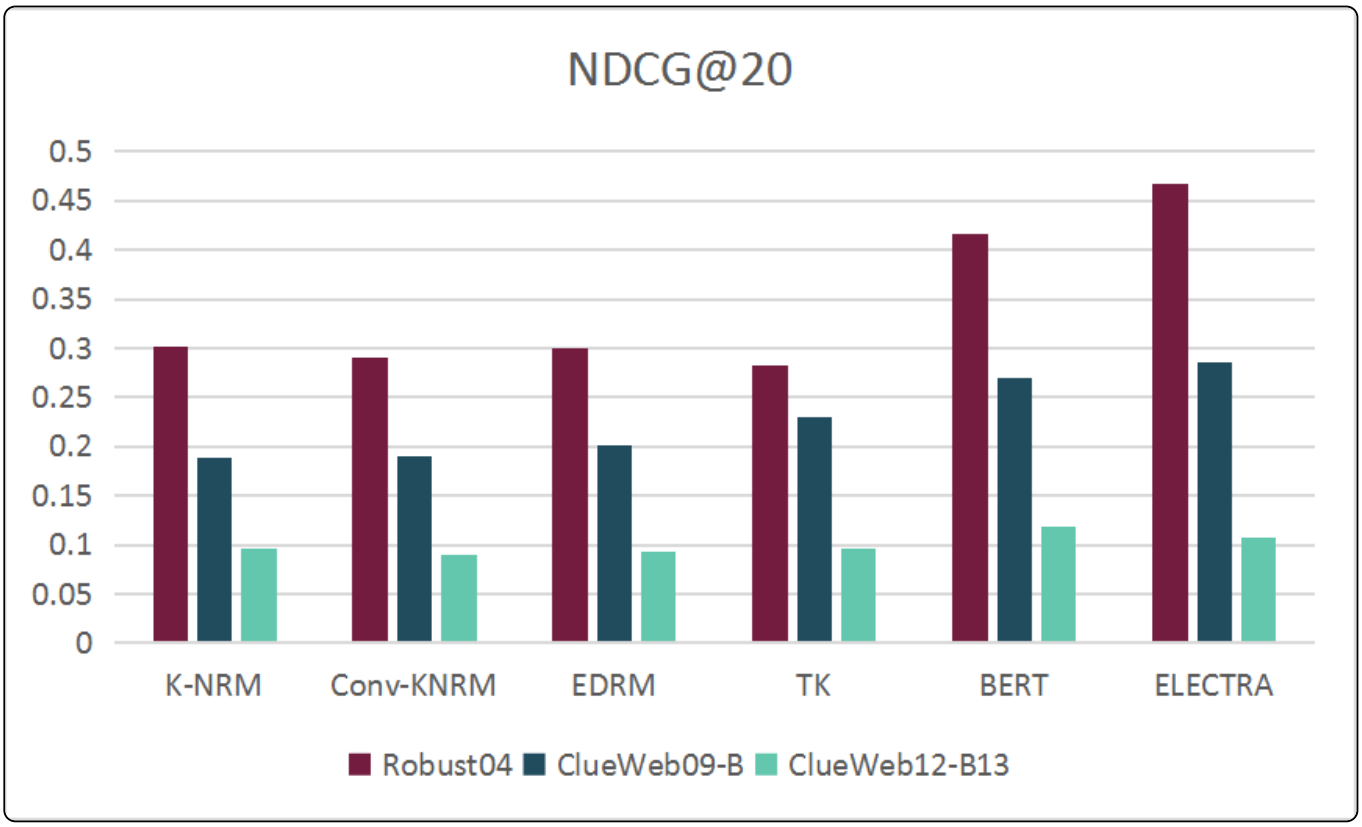


图8 OpenMatch实验结果

团队(CMT: CMU, Microsoft and THU)在近期的TREC COVID (新冠肺炎信息检索)竞赛第二轮无人工干预组的25只队伍中取得了第一名，比赛官网、相关数据、模型checkpoint及复现方法已在OpenMatch GitHub中给出。

实验结果如下图所示：

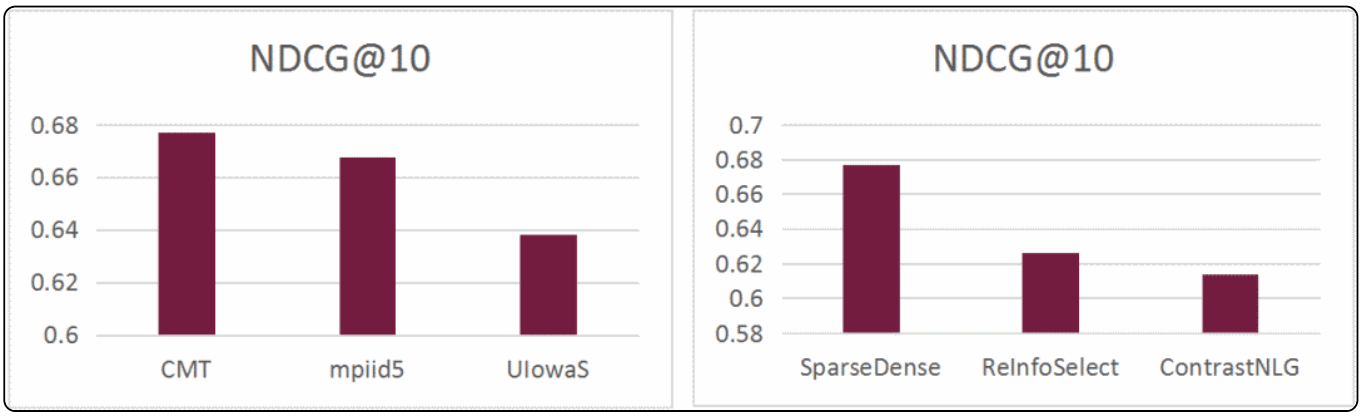


图9 新冠肺炎检索结果

左图为比赛前三名结果，右图是我们提交的三种方法：

- SparseDense：通过BM25(Sparse)+ANN(Dense)进行相关文档检索，用medical marco数据训练的SciBERT模型进行重排序。
- ReInfoSelect：通过BM25进行相关文档检索，用policy gradient筛选medical marco数据并训练SciBERT模型，进行重排序。
- ContrastNLG：通过BM25进行相关文档检索，用medical marco和生成的pseudo query[11]数据训练的SciBERT模型进行重排序。

可以看出，通过筛选数据的数据增强方法能够有效提升模型效果，通过BM25+ANN的文档检索方法可以大幅提升排序效果。另外，mpiid5团队通过BM25进行相关文档检索，使用训练的ELECTRA进行重排序，此方法也已在OpenMatch复现。

5

结语

在神经信息检索模型中，相关性标签对于训练神经网络模型来讲十分重要，然而在很多场景下，相关性标签往往是难以获得的。因此，我们通过采用融合外部知识以及引入弱监督信号的方式来增强模型的效果。我们通过开源工具OpenMatch为整个开放域信息检索提供了一套完成的解决方案以及领域迁移的工作，从而帮助神经网络模型在开放域信息检索中得到更好的效果。

相关论文

1. End-to-End Neural Ad-hoc Ranking with Kernel Pooling. Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, Russell Power. SIGIR 2017.

2. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-hoc Search. Zhuyun Dai, Chenyan Xiong, Jamie Callan, Zhiyuan Liu. WSDM 2018.

3. Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval. Zhenghao Liu, Chenyan Xiong, Maosong Sun, Zhiyuan Liu. ACL 2018.

4. SciBERT: A Pretrained Language Model for Scientific Text。Iz Beltagy, Kyle Lo, Arman Cohan. EMNLP 2019.

5. Interpretable & Time-Budget-Constrained Contextualization for Re-Ranking. Sebastian Hofstätter, Markus Zlabinger, Allan Hanbury. ECAI 2020.