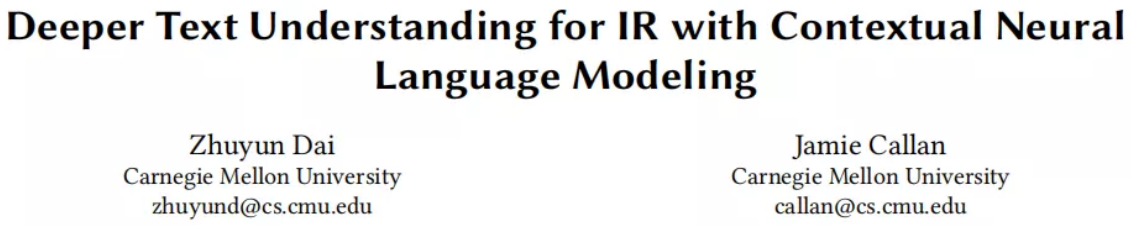
SIGIR 2019 | 使用上下文神经语言模型对IR进行更深入的文本理解

SANE 艾达AI 1周前

神经网络为自动学习复杂的语言模式和query-document关系提供了新的可能性。神经IR模型在学习查询文档相关模式方面取得了很好的效果，但在理解查询或文档的文本内容方面却鲜有探索。本文研究利用最近提出的上下文神经语言模型BERT，为IR提供更深入的文本理解。



Deeper Text Understanding for IR with Contextual Neural Language Modeling

Zhuyun Dai
Carnegie Mellon University
zhuyund@cs.cmu.edu

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

论文地址:

https://arxiv.org/pdf/1905.09217.pdf

1. 引言

文本检索需要理解文档含义和搜索任务。神经网络是一个有吸引力的解决方案，因为它们可以从原始文档文本和训练数据中获得这种理解。大多数神经IR方法都侧重于学习查询-文档相关模式，即关于搜索任务的知识。然而，只学习相关性模式需要大量的训练数据，但仍然不能很好地推广到尾部查询或新的搜索领域。这些问题使预训练的通用文本理解模型成为可取的。

本文研究利用最近提出的上下文神经语言模型BERT，探讨了BERT的语言理解对ad-hoc文档检索的影响。它验证了BERT模型在两个具有不同特性的ad-hoc检索数据集上表现。

实验结果表明，来自BERT的上下文文本表示比传统的单词嵌入更有效。 与词袋检索模型相比，上下文语言模型能够更好地利用语言结构，给使用自然语言描述的查询带来了很大的

改进。最后，通过从大型搜索日志中获取搜索知识来增强BERT，产生了一个新的预训练模型，该模型配备了关于文本理解和搜索任务的知识，这有利于在标注数据有限的情况下进行相关的搜索任务。

2. 数据集

本文使用两个具有不同特性的标准文本检索数据集。Robust04是一个拥有0.5M文档和249个查询的新闻语料库。包括两个版本的查询：简短的关键字查询（Title)和较长的自然语言查询(Description）。还包括一个说明作为相关性评估的指导。Clue Web09-B包含50M网页和200个带有Title和Description的查询。对于长文本，使用一个大小为150字，步长为75字的滑动窗口生成一个个passage。对于ClueWeb09-B，文档标题被添加到每个段落的开头。为了用搜索数据增强BERT，我们使用Bing搜索日志示例，样本包含0.1M查询和5M查询-文档对。

3. 模型

3.1 Problem Formulation

本文实验使用了现有的BERT架构，如图1所示。该模型将查询token和文档token的连接作为输入，并使用特殊token“ [SEP]”将两个段分开。token被嵌入到embedding中。为了进一步将查询与文档分开，将分段嵌入“ Q”（用于查询token）和“ D”（用于文档token）添加到token embedding中。为了捕获单词顺序，添加了位置embedding。token经过多层变压器。在每一层，通过加权所有其他token的嵌入，为每个token生成一个新的上下文化嵌入。权重由几个注意矩阵（多头注意）决定。注意度更高的单词被认为与目标单词相关性更高。不同的注意矩阵捕获不同类型的单词关系，例如完全匹配和同义词。注意力遍及查询和文档，因此要考虑查询文档的交互。最后，第一个标记的输出embedding用作整个查询文档对的表示。它被送入多层感知器（MLP）中以预测相关性的可能性（二进制分类）。该模型使用预训练的BERT权重进行初始化，以利用预训练的语言模型，而最后的MLP层是从头开始学习的。在训练期间，整个模型将进行调整以学习更多特定于IR的表示形式。

3.2 Baseline Models

无监督baseline使用Bag-Of-Words(BOW)和Sequential Dependency Model Queries(SDM)。Learning-To-Rank(LTR) baseline包括RankSVM和加入bag-of-words特征的Coor-Ascent。神经baseline包括DRMM和Conv-KNRM。

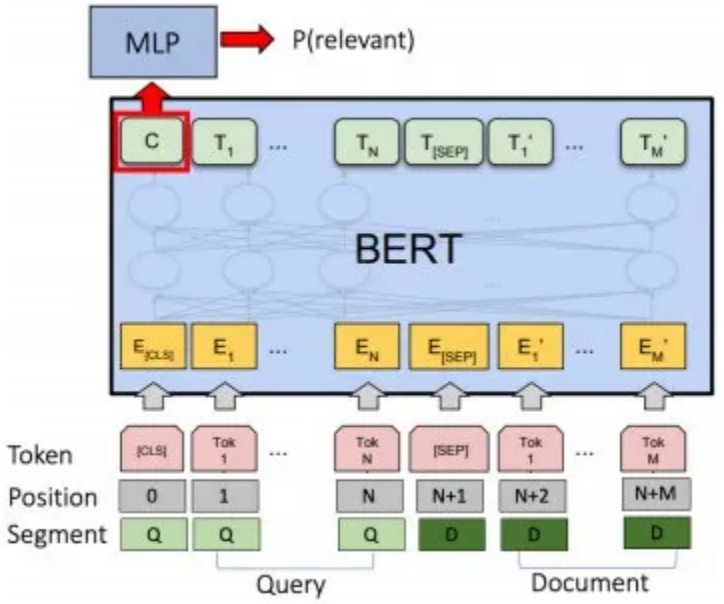


图1 BERT句子对分类架构

4. 实验

各排序方法的排序精度见表1。 在Robust04上，BERT模型始终比baseline获得更好的搜索精度，Title类型query有10%的幅度，Description类型query有20%的幅度。在Clue Web09-B上，BERT在Title类型query上与Coor-Ascent相当，在Description类型query上更好。 结果证明了BERT在文档检索中的有效性，特别是在Description类型query。

表2显示了SDM、Coor-Ascent和BERT-MaxP在Robust04上的性能。SDM对Title类型的效果最好。 Coor-Ascent在Description类型和 Narrative方面稍好一些。相比之下，BERT-MaxP通过对词义和上下文进行建模，对较长的查询进行了很大的改进。Description类型的关键字版本比SDM和Coor-Ascent的原始查询性能更好，因为stopwords对传统的匹配信号(如TF)是有噪声的。相反，BERT对原始自然语言查询更有效。虽然停止词和标点符号没有定义信息需求，但它们在语言中构建了结构。

Model	nDCG@20			
	Robust04		ClueWeb09-B	
	Title	Description	Title	Description
BOW	0.417	0.409	0.268	0.234
SDM	0.427	0.427	0.279	0.235
RankSVM	0.420	0.435	0.289	0.245
Coor-Ascent	0.427	0.441	**0.295**	0.251
DRMM	0.422	0.412	0.275	0.245
Conv-KNRM	0.416	0.406	0.270	0.242
BERT-FirstP	0.444†	0.491†	0.286	**0.272**†
BERT-MaxP	**0.469**†	**0.529**†	0.293	0.262†
BERT-SumP	0.467†	0.524†	0.289	0.261

表1 在Robust04和ClueWeb09-B上的搜索精度

Query	Avg Len	nDCG@20					
		SDM		Coor-Ascent		BERT-MaxP	
Title	3	0.427	–	0.427	–	0.469	–
Desc	14	0.404	-5%	0.422	-1%	0.529	+13%
Desc, keywords	7	0.427	-0%	0.441	+5%	0.503	+7%
Narr	40	0.278	-35%	0.424	-1%	0.487	+4%
Narr, keywords	18	0.332	-22%	0.439	+3%	0.471	+0%
Narr, positive	31	0.272	-36%	0.432	+1%	0.489	+4%

表2 不同类型Robust04查询的准确性

最后一组实验验证BERT的语言建模知识是否可以与额外的搜索知识相叠加，以建立一个更好的排序器，以及搜索知识是否可以通过领域适应的方式学习，以缓解冷启动问题。我们在Bing搜索日志示例上使用0.1M查询训练BERT，并在ClueWeb09-B上对其进行微调。结果见表3，BERT-FirstP+Bing实现了最好的性能，确认文本检索需要理解文本内容和搜索任务。BERT的简单域适应性可产生一种具有两种知识的预训练模型，可以改善标签数据受限的相关搜索任务。

Model	Knowledge		nDCG@20	
	Text	Search	Title	Desc
Coor-Ascent	Low	Low	0.295	0.251
BERT-FirstP	High	Low	0.286	0.272†
Conv-KNRM+Bing	Low	High	0.314 †	0.275 †
BERT-FirstP+Bing	High	High	**0.333**†	**0.300**†

表3 Clue Web09-B上BERT-FirstP+Bing的准确性

5. 总结

本文研究了最近提出的深度神经语言模型BERT对ad-hoc文档检索任务的影响。调整和微调BERT在两个不同的搜索任务上达到了较高的精度，打败了现在的baseline方法，显示了BERT语言模型对IR的有效性。同时，语料库训练的语言模型可以通过简单的领域适应来补充搜索知识，从而形成一个强大的排序器，能够在搜索中对文本的意义和相关性进行建模。

长按识别关注，获取更多新鲜论文

喜欢此内容的人还喜欢

Arxiv 2020 | mT5: 支持101种语言的大规模多语言预训练模型
艾达AI

傅雷家书，上流父母皆祸害
海外风云

台湾又爆"艳照门"，范玮琪老公怒怼媒体：你们连自慰都不敢讲！