arXiv:2301.04339v1 [cs.CL] 11 Jan 2023

上下文化注意力嵌入中的主题

Mozhgan Talebpour[1], Alba Garcia Seco de Herrera[1], Shoaib Jameel[2]

[1] 计算机科学与电子工程，埃塞克斯大学，英国。[2] 电子学与计算机科学，南安普顿大学，英国。 {mozhgan.talebpour, alba.garcia}@essex.ac.uk, M.S.Jameel@southampton.ac.uk

摘要。 通过预训练语言模型获得的上下文化词向量编码了多种知识，这些知识已被应用于各种任务。与这些语言模型互为补充的是概率主题模型，它们能够从文本中学习主题模式。近期研究表明，对语言模型生成的词级上下文表示进行聚类，可以模拟从潜在狄利克雷分配中发现的潜在主题词簇。一个重要的问题是：当语言模型并未被明确设计为对潜在主题建模时，这类主题词簇是如何通过聚类在语言模型中自动形成的？为了解答这一问题，我们设计了多种探测实验。通过使用 BERT 和 DistilBERT，我们发现注意力框架在建模这类词语主题簇方面发挥着关键作用。我们坚信，这项工作将为进一步研究概率主题模型与预训练语言模型之间的关系铺平道路。

1 引言

预训练语言模型（PLM），例如 ELMo [35], 生成式预训练变换器（GPT） [37], PaLM [11],以及双向 Transformer 编码器表示（BERT） [14], 使用大量文本数据进行预训练 [24], ；例如，BERT 已在 BookCorpus 和维基百科语料集上完成预训练。在领域无关预训练过程中，这些模型会编码各种潜在信息，例如语义和句法属性 [57], ；因此，即使在不同应用中的零样本设置下，这些模型也能做出可靠的预测 [20,41,43]。尽管预训练过程在计算资源 [51]和资金方面成本高昂 [47],，但这些模型可以通过低成本微调，可靠地处理各种下游任务，例如文档分类 [1]和信息检索 [61,50], ，这一过程通常称为迁移学习 [32]。例如，BERT 在自然语言理解 [63], 、文本摘要 [25], 、文档分类 [10]以及其他自然语言处理（NLP）下游应用中 [43]均展现出强大的性能。

另一类持续主导文本挖掘领域的模型是概率主题模型（PTM） [8,7]。这类模型以概率方法，在完全

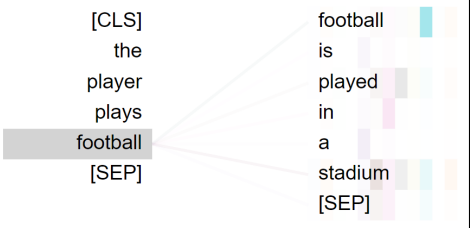
无监督的方式下确定文本语料库中的主要主题。潜在主题被描述为词语的概率分布。潜在狄利克雷分配（LDA） [8] 是一种广受欢迎的主题发现模型。在 LDA 中，模型学习将文档表示为潜在主题的混合，而每个主题则表示为词语的混合。将 LDA 视为矩阵分解模型时，给定作为输入的词项-文档共现矩阵以及主题数量，模型会将该矩阵分解为两个低维矩阵，即词-主题表示和文档主题表示。词-主题矩阵反映词汇表中的词语对各个主题的重要性，而文档-主题矩阵则反映每篇文档中的主题分布。LDA 是一种基于贝叶斯学习的热门模型，而另一类基于线性代数的模型——非负矩阵分解（NMF） [56,26] ——也同样广受欢迎，可用于学习主题 [33]。

在 [43], 中，作者剖析了 BERT，以了解每一层的特性。他们发现，较低层（即第 1 层或第 2 层）捕捉线性词序，中间层则可靠地学习句法信息，而较高层捕捉上下文信息。 [49]和 [45]中的作者展示了，通过 k 均值等简单算法对 BERT 词嵌入进行聚类，会得到类似于由主题模型学习得到的词簇。作者进行了一系列定性探测实验，发现 BERT 的大多数词簇都类似于潜在狄利克雷分配模型（LDA模型）通常发现的结果。尽管这些研究得出了相关观察，但尚未得到充分研究的是：鉴于 BERT 或其他任何上下文语言模型都不是为建模主题词簇而设计的，预训练时主题信息是如何编码的。在这项工作中，我们通过进行不同的探测实验，回答了 BERT 和 DistilBERT [44] 如何捕捉与主题模型所学习内容相似的词簇。我们发现，这些语言模型中的注意力 [4,9]机制在建模类似于主题模型所发现的词主题方面发挥着关键作用。

2 相关工作

预训练语言模型 [31]的主要目标是通过寻找最可能的词语序列和模式来模拟人类语言理解。传统语言模型使用概率分布来预测下一个词，但其可扩展性不佳，例如基于一元语法、二元语法或三元语法的语言模型 [36]。近年来发展的预训练语言模型使用大量文本数据进行训练，其中一些模型以自监督的方式采用了称为掩码语言建模的策略。这些模型完成训练后，已被应用于各种各样的领域。预训练语言模型的关键优势在于，它们能够可靠地应用于不同的下游任务 [15]。

BERT 采用堆叠的 Transformer [52]层构建，每一层都能捕捉文本数据的不同属性，例如，有些层特别适合捕捉语义信息 [53,48]。Transformer 由编码器-解码器结构组成。编码器将输入词元序列转换为高维表示。解码器根据编码器的输出预测输入数据 [18]。然而，在 BERT 中



	57	38	65	70
0	app	game	time	movie
1	apps	player	day	film
2	developer	video game	work	show
3	application	gaming	hour	story
4	user	developer	week	episode

(a) 通过第12层的可视化展示 BERT 中的注意力机制。我们观察到，(b) 从 LDA模型中获得的词语，按概率从高到低排序。那些在上下文中居于核心位置的词语被赋予了较高的注意力权重。

仅使用了 Transformer 的编码器部分。BERT 中有一个重要概念，称为注意力，它会根据底层任务中的重要性为不同的输入特征分配权重。举例来说，给定一段关于猫的文本，模型会通过注意力权重，更加关注毛发、眼睛等词语。 [12] 中的研究也探讨了 BERT 的注意力，作者发现不同的注意力头关注语言的不同方面；例如，他们发现，注意力头能够以高得多的准确率识别动词的直接宾语、名词的限定词、介词的宾语以及物主代词的宾语。虽然他们研究了不同注意力头中编码的句法和语义信息，但并未单独探究 LDA 和非负矩阵分解等主题模型所学习的潜在主题。在图1a中，我们借助一种常用的可视化工具[3]展示注意力的工作方式。我们依次输入两个句子，第一个句子“The player plays football.”之后是第二个句子“Football is played in a stadium.”，两者描述的都是足球这项运动。可视化工具展示了这样一种情况：当我们选中第一个句子中的词元“football”时，其他语义相关的词元，例如“football”、“stadium”和“played”，会以较高的注意力权重突出显示。

主题建模是一种机器学习技术，可以自动发现未标注数据中的隐藏主题。主题被定义为词语的概率分布。尽管这些主题模型的灵感源自基于潜在概念的模型，例如潜在语义分析 (LSA) [13] 和概率潜在语义分析 (pLSA) [21], ，潜在狄利克雷分配 (LDA) [8]因解决了 LSA 和 pLSA 中的一些根本性挑战（例如在大型数据集上的扩展性和过拟合问题），已被广泛用于发现潜在主题。在 [27], 中，作者证明静态词嵌入与奇异值分解 (SVD) 相关，而 SVD 是潜在语义分析中使用的核心算法；我们在此证明的是，预训练语言模型等模型会隐式地学习由预训练主题模型编码的潜在主题信息。

LDA 的训练采用了可交换性 [17]假设，即文档中的词序并不重要。这些模型将文档描述为

[3] https://github.com/jessevig/bertviz

由词语混合而成，每个文档又由用户定义的多个主题混合而成。请注意，BERT 不对文档级信息建模；Sentence-BERT（SBERT）等扩展模型 [40] 可用于对文档建模。

在图 1b 中，我们展示了使用一款免费在线主题建模可视化工具[4]通过潜在狄利克雷分配得到的典型输出。从该输出可以看出，某些主题包含五个概率排名最高的词，并以离散数字作为主题标签索引。从主题索引 57 可以推断，该主题描述的是计算机或移动应用程序及其开发。主题编号 38 描述的是电子游戏。

BERT 在许多自然语言处理下游任务中展现了最先进的成果，例如自然语言推理和信息检索。此前的一些研究强调了上下文信息作为主题建模附加特征的重要性。例如，在 [3],中，研究人员考察了句子上下文表示和神经主题模型的重要性。在 SBERT [40], 中，嵌入表示被用作 prodLDA [46]神经主题模型的输入。如果输入文档的长度超过 SBERT 预先定义的长度，则文档其余部分会被省略。尽管存在这一局限，与词袋（BoW）表示嵌入相比，该模型仍获得了更高的一致性分数。另一些研究则关注向 BERT 模型添加主题建模信息是否能够提升其性能，以及如何实现这种提升。在 Peinelt 等人开展的一项研究中，[34], 他们使用主题建模来提升 BERT 在语义相似度领域应用中的表现，例如问答。他们使用 BERT-base 最后一层的 *[CLS]* 标记嵌入，作为输入文档对应的嵌入。Wang 等人， [55] 认为通过添加主题信息可以改进 BERT 上下文嵌入。在他们的研究中，BERT 嵌入由语料库中的主题推导而来。该研究的发现表明，词向量表示等于不同主题向量的加权平均值。如果某个主题在语料库中具有较高的重要性，与该主题相关的词语也会获得更高的重要性。

在另一项相关研究中， [23], 研究人员将主题文本分类应用于一个科学领域数据集。作者将研究结果与 SciBERT [2], 进行了比较，SciBERT 是一种基于 BERT、但针对科学文档的预训练语言模型。他们将 BERT 嵌入与文档主题向量拼接，作为一个两层前馈神经网络的输入。在最近的一项研究中， [49],研究人员从不同角度考察了 BERT 嵌入的作用。该研究认为，对标记级 BERT 嵌入进行聚类，与主题建模有诸多相似之处。作者使用了不同的预训练语言模型，例如 BERT、GPT-2 [38] 和 RoBERTa 的 [28]最后三层嵌入。研究发现，除 RoBERTa 外，BERT 和 GPT-2 的词级聚类所得的簇与使用 LDA 模型获得的簇十分相似。LDA 将主题学习为词语的概率分布，而通过对预训练语言模型中的标记级嵌入进行聚类所获得的词簇

[4] https://pyldavis.readthedocs.io/en/latest/index.html

不能被误认为是词语的概率分布。作者指出，PTM 的词簇与预训练语言模型得到的簇之间存在一些相似之处。

尽管上述研究展示了预训练主题模型与预训练语言模型之间的重要关系，但目前仍缺乏对潜在主题如何编码在预训练语言模型向量中，以及哪些组成部分有助于编码这些信息的进一步理解。上述研究中，有些以统一的方式利用预训练语言模型训练潜在主题。问题在于，是否有必要再次利用预训练语言模型学习潜在主题。尽管这些研究在量化指标上取得了提升，但潜在主题如何帮助改善结果仍不清楚。

3 探测任务

本文旨在研究的问题是：潜在主题信息是否会自动编码在情境化词嵌入中。尽管潜在主题信息是否被编码并不明显，我们仍必须设计探测任务。因此，我们的主要目标是理解，BERT 和 DistilBERT 等预训练语言模型如何发现预训练主题模型通常能够发现的词簇，尽管它们并非专门为建模此类信息而设计。为此，我们首先选择更深入地研究注意力头在预训练语言模型中的作用。这是因为，与主题模型类似，处于文档全局上下文核心位置的词会被赋予较高的概率，而处于某个主题核心位置的词也会被赋予较高的概率。例如，如果文档讨论的是“体育”，那么“足球”“进球”和“球员”等词在该文档中的概率就会较高。同样，这些词在与体育相关的主题中出现的概率也会较高。注意力机制在文档中也呈现出类似的行为：在给定上下文窗口内处于核心位置的词会被赋予较高的注意力权重。注意力权重表示某个词与其他词 [12] 同时出现在预先定义的上下文窗口中时，该词的重要性。

鉴于 BERT-base uncased 和 DistilBERT-based uncased 模型广受欢迎且计算简便，我们将其作为预训练语言模型。我们还知道，LDA模型会输出词主题和文档主题表示 [8]。给定因子或潜在维度的数量，非负矩阵分解会将共现矩阵分解为两个低维矩阵，其中一个矩阵编码词簇，另一个矩阵编码文档簇。由于语言模型能够捕捉词级模式，因此我们在 LDA 和非负矩阵分解中选择词主题。由于 LDA 和非负矩阵分解都可以根据其概率值明确地分配到软聚类中，因此对于注意力表示，我们必须使用软聚类算法对其进行聚类。这将有助于我们生成带有簇分配的词簇。

我们还可以研究其他组成部分，例如不同 Transformer 层堆叠在一起时所发挥的作用。然而，先前的研究已经发现，不同层会捕捉文本数据的不同特

性。例如，在 BERT 中，较低层捕捉线性词序，中间层捕捉句法信息，而较高层捕捉语义信息。经过充分实验，这些研究均未发现其中某一层还会对类似潜在主题的词簇进行建模。因此，鉴于这些研究的发现，我们首先关注预训练语言模型中的注意力头。

在 BERT-base 中，共有 12 层，每层包含 12 个注意力头。注意力头会计算输入句子中所有词语组合之间的注意力权重。注意力权重可以看作是同时考量两个词语时的一项重要指标。例如，(weather, sunny) 这一词对的注意力权重高于 (weather, desk) 这一词对。这是因为 BERT 在数十亿个词元上进行训练时，(weather, sunny) 组合出现的频率高于与“desk”等其他词语组成的组合。同样，在潜在狄利克雷分配模型中，如果出现“sports”和“football”等词语，它们在词主题中的概率值就会较高。DistilBERT 同样基于 BERT-base 模型，但参数量大幅减少，因此模型更加轻量。它是通过一种称为知识蒸馏的过程得到的 [29,19]： 使用被称为教师模型的原始大型模型训练经过压缩的轻量学生模型，使其模仿教师模型的行为。研究发现，DistilBERT 在大幅减少参数量的同时，保留了 BERT 的大部分优势。

我们使用两个公开可用的基准数据集，开展两项不同的探测任务，以证明研究发现具有普遍适用性。在第一项探测任务中，我们对从 PTM 和 PLM 模型中获得的表示进行词级聚类，并计算一致性度量；该度量常用于主题模型，以评估主题的质量。对于语言模型，我们从模型的每一层提取注意力权重，并为词汇表中的每个词获取词注意力表示。随后，我们使用聚类算法对这些注意力向量进行聚类，并将注意力向量作为特征。通过这种注意力聚类，我们希望语义相关的词能够被归入同一个簇。其动机在于，如果词簇中包含主题相关的词，这些簇就会呈现较高的一致性度量。尽管一致性度量的实用性一直存在争议，[22],在本研究中，我们仍采用相同的度量对所有模型进行定量比较。

我们旨在探究，预训练语言模型的一层与从预训练主题模型中获得的词主题表示之间是否具有可比的一致性表现。所谓可比，是指二者的一致性结果在数值上是否接近。如果一致性结果具有可比性，那么我们可以认为，就词语的主题建模而言，语言模型和主题模型正在学习语义相关的内容。尽管不能完全依赖一致性探测任务，我们还设计了一个额外的探测任务，用于找出从预训练语言模型和预训练主题模型中获得的词簇之间的词重叠。我们的动机是，如果簇之间的一致性值较高，那么簇中的词之间必然存在可靠的重叠。由于在 BERT-base 模型中，第 10、11 和 12 层比低层捕获了更多的语义

信息，我们预计，语言模型高层中的词簇将与预训练语言模型所学习的那些词簇表现出更高的共同性。

3.1 实验设置

数据集：我们使用了 20个新闻组（20NG）和 IMDB 数据集，这两者都是文本挖掘领域常用的热门数据集。移除重复和空实例后，20NG 数据集包含约 18,000 篇文档，分属 20 个新闻类别。IMDB 数据集包含 50,000 条电影评论，标注为正面或负面。20NG 数据集包含一些较长的文档，而 IMDB 数据集中的文档则相对较短，且文本噪声相对较多。

文本预处理：对于预训练主题模型，我们采用了常见的预处理策略，例如移除停用词、标点符号和非 ASCII 字符。通过实验，我们发现，如果不从文本中移除停用词，它们往往会主导大多数主题，还会增加语义空间的维度，导致较高的空间和时间复杂度。尽管已有研究提出了一些使用预训练主题模型对自然语言建模的变通方法，例如使用非对称先验，但这些方法在大型数据集上可能计算开销很大 [54]。对于预训练语言模型，我们让默认的预处理器处理预处理工作，例如，BERT-base 模型使用 WordPiece 分词器。我们使用 NLTK [5], 进行了句子切分。

预训练语言模型注意力权重：对于词汇表中的每个词，我们从 BERT-base uncased 和 DistilBERT 模型中获取该词的注意力权重。由于 BERT 使用 WordPiece 分词，如果分词后的句子长度超过 512 个词元，就会拆分输入句子；这在相关文献中很常见。句子中所有词元的注意力权重都会被存储。如果一个词出现在不同的句子中，则使用该词所有注意力的平均值；这也是一种常见做法，包括对其各个词片段的嵌入取平均值 [59]。我们获取了 BERT 每一层的注意力权重。BERT 的注意力权重定义为每一层中所有注意力头的平均值。

我们从原始 BERT-base 模型中获取了注意力权重。此外，我们还从 BERT 模型的微调版本中获取了注意力权重，以评估微调在这一过程中可能发挥的作用。微调是在 20NG 和 IMDB 数据集上进行的，使用与已标注实例相关联的标签完成文本分类任务。通过微调过程中的交叉验证，我们展示了表现最佳模型在测试集上的结果；该模型的理想参数是通过 30% 的留出数据集获得的。对于原始 DistilBERT-base 模型，我们采用了相同的配置。

主题建模：我们使用 Gensim 实现的潜在狄利克雷分配（LDA）模型，以在我们的数据集中发现潜在主题。在 20NG 数据集中，我们将主题数量设定在 20 至 200 之间。在 IMDB 数据集中，我们将主题数量设定在 2 至 30 之间，这带来了更好的

20个新闻组			
LDA		NMF	
# 主题	c-v	# 主题	c-v
20	**0.518**	20	0.478
30	0.487	30	**0.504**
50	0.504	50	0.484
100	0.474	100	0.453
150	0.470	150	0.455
200	0.473	200	0.474

IMDB			
LDA		NMF	
# 主题	c-v	# 主题	c-v
2	0.363	2	0.276
5	0.364	5	0.275
10	0.370	10	0.299
20	**0.461**	20	**0.300**
30	0.437	30	0.299

表1：潜在狄利克雷分配和非负矩阵分解模型的一致性结果。

结果。我们使用了 Gensim 实现的非负矩阵分解模型。根据[58], ，较大的数据集往往比规模较小的数据集包含更多主题。因此，我们为不同的数据集选择了不同的主题数量范围。我们没有将主题数量设定为等于从预训练主题模型（PTMs）中获得的词向量的维度，因为预训练主题模型（PTMs）往往会在其向量中编码多种信息，例如句法信息和语义信息。此外，主题数量大幅超过上述选择的数量，往往会产生次优的潜在主题，从而导致性能下降。

聚类：我们对从预训练语言模型获得的嵌入使用了软高斯混合模型（GMM）[6] 聚类算法。潜在狄利克雷分配本身就是一种软聚类模型，它使用概率值为实例分配软聚类 [60]。在潜在狄利克雷分配中，我们可以根据每个主题中词语的概率值自动获得词-主题分配；对于通过高斯混合模型获得的簇，这一点同样成立。我们使用高斯混合模型，是因为它的实现广泛可用，并且可通过不同的软件库免费获取。

评估：在主题建模中，一致性度量已被广泛用于评估潜在主题的质量 [30]。我们采用了一致性分数“c-v”，该分数可通过 Gensim 库获取。这一度量改编自 Roder 等人的研究 [42]。我们使用一致性来衡量从预训练语言模型和预训练主题模型中获得的词簇内词元的语义相关性。我们还统计两个模型所得簇中前 k 个词语之间的词重叠数量，以衡量各簇之间的词重叠。我们设置了 $k = 20$ ，这在选取主题相关性最高的前 k 个词语，同时避免选择词簇中估计概率较低的（通用或噪声）词语之间，取得了可靠的折中。为计算词重叠值，对于预训练主题模型中的每个主题以及预训练语言模型中每一层的词簇，我们计算了前 k 个词语之间的重叠，随后求取其“众数”值。虽然存在熵和排他性等指标 [49]，我们将在本文的扩展版本中使用这些指标。

4 讨论

我们已在两个不同的数据集上计算了簇的一致性值。给定两个簇及各自的一致性值，如果一个簇的一致性值高于另一个，则一致性值较高的簇被视为

层	VB30	VB50	VB100	VB150	VB200	微调30	微调50	微调100	微调150	微调200
1	0.360	0.502	0.489	0.481	0.343	0.333	0.477	0.502	0.466	0.330
2	0.346	0.480	0.463	0.464	0.334	0.327	0.479	0.480	0.462	0.333
3	0.329	0.450	0.453	0.448	0.323	0.315	0.466	0.450	0.459	0.324
4	0.328	0.466	0.461	0.452	0.332	0.324	0.466	0.466	0.461	0.332
5	0.33	0.460	0.448	0.449	0.324	0.325	0.458	0.460	0.453	0.329
6	0.33	0.459	0.455	0.451	0.318	0.347	0.466	0.459	0.460	0.337
7	0.337	0.478	0.471	0.454	0.325	0.346	0.495	0.478	0.479	0.347
8	0.347	0.469	0.468	0.470	0.336	0.353	**0.508**	0.469	**0.496**	0.359
9	0.346	0.486	0.474	0.471	0.344	**0.370**	**0.508**	0.486	0.494	0.360
10	0.373	0.480	0.483	0.476	0.360	0.368	0.502	0.480	0.494	0.358
11	0.369	**0.503**	0.489	0.481	0.360	0.357	0.483	**0.503**	0.489	**0.361**
12	**0.373**	0.502	**0.489**	**0.484**	**0.363**	0.364	0.485	0.502	0.480	0.355

层	VB2	VB5	VB10	VB20	VB30	微调2	微调5	微调10	微调20	微调30
1	0.411	0.390	0.365	0.358	0.355	0.455	0.442	0.372	0.348	0.333
2	0.473	0.447	0.374	0.347	0.352	0.469	**0.459**	0.420	0.391	0.384
3	0.480	0.414	0.418	0.385	0.366	0.501	0.452	0.415	**0.412**	**0.404**
4	**0.586**	0.478	0.444	0.421	0.404	0.457	0.388	0.386	0.383	0.380
5	0.583	**0.490**	0.439	0.426	0.422	0.410	0.383	0.350	0.359	0.356
6	0.563	0.477	**0.471**	0.429	0.405	0.452	0.438	0.396	0.374	0.357
7	0.546	0.485	0.431	0.425	0.416	0.489	0.403	0.399	0.374	0.366
8	0.510	0.438	0.428	0.414	0.415	**0.514**	0.427	0.395	0.397	0.369
9	0.452	0.410	0.393	0.383	0.373	0.438	0.425	0.366	0.377	0.374
10	0.476	0.430	0.381	0.351	0.349	0.426	0.417	0.369	0.361	0.349
11	0.425	0.429	0.400	0.398	0.385	0.430	0.391	0.346	0.354	0.347
12	0.523	0.454	0.446	**0.439**	**0.431**	0.469	0.433	**0.440**	0.402	0.385

表2：20NG 高斯混合模型（左）和 IMDB 高斯混合模型（右）基于左侧和右侧的 BERT-base 注意力权重进行聚类。数值表示一致性结果。VB 指原始 BERT-base 模型，FT 指微调版本。VB 和 FT 后面的数字表示高斯混合模型中指定的簇数。

层	VD30	VD50	VD100	VD150	VD200	微调30	微调50	微调100	微调150	微调200
1	0.504	0.497	0.518	0.515	0.521	0.502	0.508	0.511	0.517	0.513
2	0.509	0.509	0.510	0.514	0.515	0.511	0.518	0.510	0.507	0.508
3	0.514	0.514	0.508	0.508	0.510	0.507	0.503	0.503	0.499	0.507
4	0.516	0.509	0.516	0.517	0.513	0.502	0.502	0.504	0.503	0.505
5	0.548	0.550	0.551	0.544	0.549	0.544	0.550	0.546	0.543	0.544
6	**0.593**	**0.572**	**0.572**	**0.573**	**0.568**	**0.573**	**0.576**	**0.573**	**0.571**	**0.571**

层	VD2	VD5	VD10	VD20	VD30	微调2	微调5	微调10	微调20	微调30
1	0.231	0.258	0.249	0.250	0.251	0.219	0.224	0.226	0.248	0.238
2	0.166	0.211	0.212	0.224	0.230	0.255	0.225	0.231	0.232	0.238
3	0.231	0.244	0.235	0.251	0.244	0.253	0.225	0.230	0.235	0.234
4	0.151	0.228	0.204	0.228	0.234	0.170	0.223	0.218	0.219	0.217
5	0.317	**0.347**	0.307	0.270	0.264	0.252	0.268	**0.274**	**0.272**	0.261
6	**0.334**	0.327	**0.325**	**0.317**	**0.312**	**0.288**	**0.270**	0.267	0.254	**0.264**

表3：20NG 上的 GMM（左）和 IMDB 上的 GMM（右）对 DistilBERT 注意力权重进行的聚类。数值表示一致性结果。VD 指原始 DistilBERT 模型，FT 指微调版本。VD 和 FT 后面的数字表示高斯混合模型指定的簇数量。

例如，在文本中，连贯簇中的词元往往在语义上彼此相关。对于潜在狄利克雷分配和非负矩阵分解模型，我们改变了主题数量，以展示主题簇的影响。表1展示了潜在狄利克雷分配和非负矩阵分解模型在20个新闻组和IMDB数据集上的主题连贯性结果。我们观察到，在潜在狄利克雷分配模型中，当主题数量为20时，20NG 数据集上的一致性值最高，为0.518。对于非负矩阵分解模型，当因子数量为30时，20NG 数据集上的一致性值最高，为0.504。在 IMDB 数据集中，潜在狄利克雷分配模型在主题数量为20时同样取得最高的一致性值，为0.461；而非负矩阵分解模型在因子数量为20时的数值为0.300。

预训练语言模型与 *20NG* 数据集：在表2（左）中的原始 BERT-base 模型（即 20NG 数据集）这一情形下，我们注意到，当软注意力簇的数量为50时，其一致性结果表现相当。具体而言，从表中可以看出，VB50 在第11层的一致性值为0.503。这一数值与主题数为20时的0.518数值接近；而在非负矩阵分解模型中，因子数为30时的一致性值约为0.504。这表明，当软聚类数量为50时，潜在狄利克雷分配和原始 BERT-base 模型的注意力词簇在语义上都具有较高的一致性。我们还注意到，上下文层主要发挥着

关键作用，能够对这类语义相近的词进行建模，即第11层。查看表4中的词重叠情况时，我们发现，前20个重叠词在 BERT-base 模型的第7、8、9和11层中也一致。这意味着20个词中有17个是重叠词。

比较在分类任务上进行微调的 BERT-base 模型微调版本的结果后，我们注意到，表2中的软簇50和100取得了与表1中潜在狄利克雷分配和非负矩阵分解模型相当的一致性表现。具体而言，从表中可以看到，当簇数为50和100时，我们得到的一致性值分别为0.508和0.503，这在数值上同样与一致性表中潜在狄利克雷分配模型的0.518以及非负矩阵分解模型的0.504相当，即表1中的结果。虽然理想情况下这些一致性结果应当相等，但考虑到数据中的噪声以及启动这些语义模型训练过程时涉及的随机性，要获得这样的结果并不容易。
BERT-base 模型微调版本的一个有趣之处在于，有两层表现出相当的一致性，并且这两层都能学习上下文信息。

考察 20NG 数据集中与“计算技术”相关的主题时，我们注意到，“organisation”“com”和“nntp”等词属于重叠词，这表明 BERT 和潜在狄利克雷分配在主题上学到了相同的词。尽管有人可能会认为，即使是 k 均值这类简单的聚类算法，也能生成连贯且重叠词较多的簇，但我们发现 k 均值并不能生成连贯簇，而且词语重叠数量也很少。例如，在大多数情况下，我们发现词语重叠值有时为 1，而更多时候为 0。

对于表3（左）所示的原始 DistilBERT 模型，我们注意到，较高层的软聚类一致性结果最高。我们发现，与表1中的非负矩阵分解模型相比，上下文层的簇一致性更高，与潜在狄利克雷分配模型的表现更为接近。例如，具有 200 个软聚类的原始 DistilBERT 版本，其表现与表1中的潜在狄利克雷分配模型相对接近。有人可能会指出，就绝对数值而言，如果只看 DistilBERT 最高层的数值，表3中的结果远高于表1。原因之一是两个模型采用了不同的预处理策略。不过，这种差异不可避免，因为在预训练主题模型中纳入停用词会导致主题含有噪声。请注意，其他层的表现也不错，例如原始 DistilBERT 模型中的第4层、30个软聚类，其结果与潜在狄利克雷分配的一致性结果相当。对于非负矩阵分解模型，DistilBERT 模型中的第4层与30个软聚类相比，表现也较为稳定。

预训练语言模型与 *IMDB* 数据集：在 IMDB 数据集中，表1展示了 LDA 和非负矩阵分解模型的主题/因子数量为 20 时的理想一致性值。LDA模型的一致性值为 0.461，非负矩阵分解模型的一致性值为 0.300。参见表2（右侧），我们可以看到

当软聚类数量为 10 时，原始 BERT-base 版本的第 6 层得到了与 LDA 相当的值。在微调版本中，与 LDA模型相比，当软聚类数量为 150 时，第 8 层的值与之相当。如果我们考察表1中的主题 30，会发现表2中第 12 层有两个相当的值；该层比其他任何层都能捕捉更多上下文信息，且对应原始版本中软聚类数量为 20 和 30 的情况。

在表4中，大部分词重叠出现在第5、9、11和12层，这些结果与20NG数据集的结果一致：较高的上下文层具有最大的词重叠。我们还注意到，第6层及以上的层具有最理想的一致性值，这表明如果这些簇具有较好的一致性，它们也会有最大的词重叠。这意味着这些簇共享一些常见词。在 DistilBERT 中，从表3和表4可以看到，非负矩阵分解模型在较高层往往呈现出相近的一致性值。在表4中，我们观察到词重叠较为均匀地分布在各层。虽然较低层显示出最大的词重叠，但我们也注意到较高层的词重叠情况相似。不过，它们的一致性值并不相当。这是因为 IMDB 样本是简短且带有噪声的句子，与20NG数据集不同，模型似乎无法十分可靠地处理这些样本。结果中另一个值得注意之处是，与非负矩阵分解模型相比，微调后的 DistilBERT 模型的一致性表现并不相当。这可能表明，分类微调会使 DistilBERT 丢失潜在主题信息。

总结如下：1）注意力机制是预训练语言模型中的重要组成部分，有助于捕捉一些预训练主题模型也能捕捉到的模式。2）预训练语言模型与预训练主题模型所得的一致性结果之间存在对应关系，因为在大多数情况下，两者的一致性表现相当。3）在预训练语言模型中，上下文化层与预训练主题模型得到的词簇之间存在较高的词重叠。4）在大多数情况下，与预训练主题模型最具共性的，是上下文化层。

我们工作的局限之一是，没有针对与 BERT 截然不同的其他语言模型（如 XLNet [62]和 GPT-3 [16]）进行实验，以确认能否从这些模型中得出类似结论。不过，值得注意的是，我们的结论指向了注意力机制的重要性，而非预训练的方式、用于预训练模型的数据集大小或模型设计。我们还需要验证这些结果是否也适用于更大型的模型，例如 BERT-large；研究此类模型需要更多的计算资源。

我们通过图2展示另一项发现，说明注意力机制的重要性，以及主题权重（概率）和注意力权重在给定上下文中往往会关注相同的词语。为生成该图，我们从 IMDB 数据集中选取了一个示例。在 BERT-base 模型中，我们考察第11层，因为它是上下文层，且在表4中具有最高的词重叠率。对于 DistilBERT-base 模型，我们

层	20NG	IMDB
1	16	14
2	16	13
3	16	12
4	16	14
5	16	**17**
6	16	14
7	**17**	12
8	**17**	12
9	**17**	**17**
10	16	11
11	**17**	**17**
12	16	**17**

层	20NG	IMDB
1	**12**	**10**
2	**12**	**10**
3	9	**10**
4	**12**	**10**
5	**12**	**10**
6	**12**	9

表4：BERT（左）和 DistilBERT（右）的注意力词语与潜在狄利克雷分配的重叠。

	words	BERT-based	DistilBERT-based	LDA	NMF
0	This	0.000000	0.000000	0.000000	0.000000
1	movie	0.071700	0.091100	0.041000	0.000000
2	is	0.000000	0.000000	0.000000	0.000000
3	terrible	0.111100	0.086300	0.002000	0.001000
4	but	0.000000	0.000000	0.000000	0.000000
5	it	0.000000	0.000000	0.000000	0.000000
6	has	0.000000	0.000000	0.000000	0.000000
7	some	0.000000	0.000000	0.000000	0.000000
8	good	0.102200	0.071100	0.010000	0.101000
9	effects	0.111700	0.097400	0.002000	0.003000

图2：以 IMDB 数据集中的一个句子为例，展示注意力。我们展示了 BERT-base 第11层和基于 DistilBERT 的第5层的结果。预训练主题模型（PTM）的主题/因子数量为 20。该图用于说明，这些模型倾向于关注上下文中的相关词元，并为停用词等通用词元分配较低的权重。

选择了第5层，因为它是上下文层之一，且在表4中具有较高的词重叠率。根据表1的结果，我们将主题数量和非负矩阵分解因子数量均设为20。从图中可以看到，所有模型都倾向于关注上下文中的相关关键词。例如，我们发现预训练语言模型会关注“good”“effects”“terrible”“movie”等与电影相关的词语，而预训练主题模型在此上下文中也倾向于关注相同的词元。从图中我们了解到，尽管预训练主题模型和预训练语言模型有所不同，但二者都倾向于关注给定上下文窗口中的相关词语。该图有助于我们探究注意力权重与主题概率之间的某些关系，因为它们都只关注重要词语。我们还注意到，模型赋予停用词等常见词语的权重较低。

虽然 [49] 中的作者发现，从某些预训练语言模型中得到的词簇，往往会将上下文化词向量聚在一起，而这些词向量与主题模型所学习到的内容相似，但我们的研究结果表明，起关键作用的是注意力机制；揭示这一点正是我们工作的主要贡献。我们也可以认为，从预训练语言模型获得的上下文化词元嵌入能得出几乎相似的结论；但在本研究中，我们希望明确探究注意力权重的作用。

5 结论

过去十年间，主题建模一直是主流的建模范式，文献中已提出多种主题模型 [64]。主题模型不仅采用贝叶斯统计进行建模，也采用基于线性代数的方法，例如非负矩阵分解模型。尽管这两类模型的建模方式不同，但它们往往呈现相似的聚类特性。随着预训练语言模型的发展，这些模型如今已在文本挖掘和自然语言处理领域占据主导地位，因为它们的表现超越了现有基线。近期研究指出，对BERT嵌入进行词级聚类，得到的词簇与使用主题模型发现的词簇密切相关。因此，我们希望探究语言模型中的哪个组成部分有助于捕捉此类主题信息，尤其是在模型并未专门用于建模潜在词主题的情况下。通过探测任务，我们发现，注意力机制在建模词模式方面发挥着关键作用，而这些词模式与使用主题模型发现的模式相似。我们坚信，本研究有助于进一步揭示主题模型与预训练语言模型之间的关系，包括注意力机制在语言模型中发挥的作用。未来，我们将开展全面的理论分析，探究主题模型与预训练语言模型之间关键的理论相似之处。我们还将研究基于 BERT 以外架构的不同预训练语言模型如何利用注意力权重编码潜在主题。

我们的研究结果不仅适用于自然语言处理和文档建模领域，也与信息检索密切相关。例如，在信息检索场景中，我们可以只使用从预训练语言模型中获取的特征来检索相关文档，而不必考虑潜在主题特征；这些特征可能会增加特征数量，甚至降低信息检索引擎的性能。此外，我们还可能向信息检索模型注入更多冗余特征。研究表明，主题模型能够提升信息检索效果，而预训练语言模型展现出的效果甚至更好。这可能是因为预训练语言模型已经在其丰富的向量空间中编码了各种特征，其中也包括潜在主题。因此，我们观察到的提升也源于预训练语言模型注意力向量中隐式编码的主题。由此，我们相信本文也将对信息检索领域产生重大影响。

参考文献

1. Adhikari, A., Ram, A., Tang, R., Lin, J.: Docbert：用于文档分类的 Bert。arXiv 预印本 arXiv:1904.08398（2019） 2. Beltagy, I., Lo, K., Cohan, A.: SciBERT：面向科学文本的预训练语言模型。arXiv（2019） 3. Bianchi, F., Terragni, S., Hovy, D.: 预训练是热门主题：上下文化文档嵌入提升主题连贯性。arXiv（2020） 4. Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., Watrin, P.: 注意力是否具有解释性？关于这场争论的导论。载于：计算语言学协会第60届年会论文集（第1卷：长篇论文）。第3889–3900页。计算语言学协会，爱尔兰都柏林（2022年5月）。https://doi.org/10.18653/v1/2022.acl-long.269, https: //aclanthology.org/2022 269 5 . acl-long. . Bird, S., Klein, E., Loper, E.: 使用 Python 进行自然语言处理：借助自然语言工具包分析文本。“O’Reilly Media, Inc.”（2009） 6. Bishop, C.M.: 模式识别。机器学习**128**(9)（2006） 7. Blei, D.M.: 概率主题模型。《ACM 通讯》**55**(4), 77–84（2012） 8. Blei, D.M., Ng, A.Y., Jordan, M.I.: 潜在狄利克雷分配。JMLR**3**, 993–1022（2003） 9. Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., Wattenhofer, R.: 关于 Transformer 的可识别性。arXiv 预印本 arXiv:1908.04211 （ 2019 ） 10. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Androutsopoulos, I.: 欧盟立法的大规模多标签文本分类。arXiv（2019） 11. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., 等：PaLM：借助 Pathways 扩展语言建模。arXiv 预印本 arXiv:2204.02311（2022） 12. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: BERT 关注什么？对 BERT 注意力的分析。arXiv（2019） 13. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: 基于潜在语义分析的索引编制。《美国信息科学学会会刊》**41**(6), 391–407（1990） 14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT：面向语言理解的深层双向 Transformer 预训练。arXiv（2018） 15. Edunov, S., Baevski, A., Auli, M.: 用于语言生成的预训练语言模型表示。arXiv（2019） 16. Floridi, L., Chiriatti, M.: GPT-3：其性质、范围、局限与后果。《心灵与机器》**30**(4), 681–694（2020） 17. Foti, N.J., Williamson, S.A.: 贝叶斯非参数模型中不可交换先验综述。《IEEE 模式分析与机器智能汇刊》**37**(2), 359–371（2013） 18. Futami, H., Inaguma, H., Ueno, S., Mimura, M., Sakai, S., Kawahara, T.: 为序列到序列自动语音识别蒸馏 BERT 的知识。arXiv（2020） 19. Hahn, S., Choi, H.: 自然语言处理中的自知识蒸馏。arXiv 预印本 arXiv:1908.01851（2019） 20. Heo, S.H., Lee, W., Lee, J.H.: MCBERT：将动量对比学习与 BERT 结合用于零样本槽位填充。arXiv 预印本 arXiv:2203.12940（2022） 21. Hofmann, T.: 概率潜在语义分析。arXiv（2013）

22. Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., Resnik, P.：自动主题模型评估是否失效？主题连贯性并不连贯。《神经信息处理系统进展》**34**，2018–2033（2021） 23. Iida, F., Pfeifer, R., Steels, L., Kuniyoshi, Y.：人工智能讲义（计算机科学讲义子系列）：序言。人工智能**3139**（2004） 24. Lai, Y.A., Lalwani, G., Zhang, Y.：预训练掩码语言模型的上下文分析。载于：EMNLP。第3789–3804页（2020） 25. Lamsiyah, S., Mahdaouy, A.E., Ouatik, S.E.A., Espinasse, B.：一种基于从BERT多任务微调迁移学习的无监督抽取式多文档摘要方法。JIS（2021） 26. Lee, D.D., Seung, H.S.：通过非负矩阵分解学习物体的组成部分。《自然》**401**（6755），788–791（1999） 27. Levy, O., Goldberg, Y.：作为隐式矩阵分解的神经词嵌入。《神经信息处理系统进展》**27**（2014） 28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.：RoBERTa：一种经过稳健优化的BERT预训练方法。arXiv（2019） 29. Lopes, R.G., Fenu, S., Starner, T.：深度神经网络的无数据知识蒸馏。arXiv预印本 arXiv:1710.07535（2017） 30. Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A.：优化主题模型中的语义连贯性。载于：2011年自然语言处理实证方法会议论文集。第262–272页（2011） 31. Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heinz, I., Roth, D.：大型预训练语言模型推动自然语言处理的最新进展：综述。arXiv预印本 arXiv:2111.01243（2021） 32. Mozafari, M., Farahbakhsh, R., Crespi, N.：一种用于在线社交媒体仇恨言论检测的基于BERT的迁移学习方法。载于：ICCNA。第928–940页。Springer（2019） 33. de Paulo Faleiros, T., de Andrade Lopes, A.：关于非负矩阵分解算法与潜在狄利克雷分配之间的等价性。载于：ESANN（2016） 34. Peinelt, N., Nguyen, D., Liakata, M.：TBERT：主题模型与BERT强强联手，实现语义相似度检测。载于：ACL。第7047–7055页（2020） 35. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettle- moyer, L.：深度情境化词表示。CoRR **abs/1802.05365**（2018） 36. Ponte, J.M., Croft, W.B.：信息检索的语言建模方法。载于：ACM SIGIR论坛。第51卷，第202–208页。美国计算机协会，纽约州纽约（2017） 37. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.：通过生成式预训练提升语言理解能力（2018） 38. Radford, A., Wu, J., Child, R., Luan, D., A modei, D., Sutskever, I., 等：语言模型是无监督的多任务学习者。OpenAI博客**1**（8），9（2019） 39. Rehurek, R., Sojka, P.：Gensim——用于向量空间建模的Python框架。NLP中心，信息学院，马萨里克大学，布尔诺，捷克共和国**3**（2）（2011） 40. Reimers, N., Gurevych, I.：Sentence-BERT：使用孪生BERT网络生成句子嵌入。arXiv（2019） 41. Ristoski, P., Lin, Z., Zhou, Q.：KG-ZESHEL：知识图谱增强的零样本实体链接。载于：第11届知识捕获会议论文集。第49–56页（2021）

42. Röder, M., Both, A., Hinneburg, A.：探索主题连贯性度量的空间。载于：第八届美国计算机协会网络搜索与数据挖掘国际会议论文集。第399–408页（2015） 43. Rogers, A., Kovaleva, O., Rumshisky, A.：BERT学入门：我们对BERT工作原理的了解。计算语言学协会汇刊**8**，842–866（2020） 44. Sanh, V., Debut, L., Chaumond, J., Wolf, T.：DistilBERT，即BERT的蒸馏版本：更小、更快、更便宜、更轻量。arXiv预印本 arXiv:1910.01108（2019） 45. Sia, S., Dalmia, A., Mielke, S.J.：厌倦主题模型了吗？预训练词嵌入的簇也能快速生成优质主题！arXiv（2020） 46. Srivastava, A., Sutton, C.：用于主题模型的自编码变分推断。arXiv（2017） 47. Strubell, E., Ganesh, A., McCallum, A.：深度学习在自然语言处理中的能耗与政策考量。arXiv预印本 arXiv:1906.02243（2019） 48. Tenney, I., Das, D., Pavlick, E.：BERT重新发现了经典自然语言处理流程。arXiv预印本 arXiv:1905.05950（2019） 49. Thompson, L., Mimno, D.：利用上下文化词表示聚类进行主题建模。arXiv（2020） 50. Trabelsi, M., Chen, Z., Davison, B.D., Heflin, J.：用于文档检索的神经排序模型。《信息检索期刊》**24**（6），400–444（2021） 51. Turc, I., Chang, M.W., Lee, K., Toutanova, K.：博览群书的学生学得更好：预训练紧凑模型的重要性。arXiv预印本 arXiv:1908.08962（2019） 52. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.：Attention Is All You Need。神经信息处理系统大会 **30** （2017） 53. Voita, E., Sennrich, R., Titov, I.：Transformer中表示的自底向上演化：机器翻译与语言建模目标下的研究。arXiv预印本 arXiv:1909.01380（2019） 54. Wallach, H., Mimno, D., McCallum, A.：重新思考潜在狄利克雷分配：先验为何重要。《神经信息处理系统进展》**22** （2009） 55. Wang, Y., Bouraoui, Z., Anke, L.E., Schockaert, S.：利用主题感知提及选择，从上下文化语言模型中推导词向量。arXiv（2021） 56. Wang, Y.X., Zhang, Y.J.：非负矩阵分解：全面综述。《IEEE知识与数据工程汇刊》**25**（6），1336–1353（2012） 57. Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., 等：探究BERT的语言知识：五种针对否定极性项的分析方法。arXiv预印本 arXiv:1909.02597（2019） 58. Wei, X., Croft, W.B.：基于潜在狄利克雷分配的文档模型用于即席检索。载于：信息检索研究与发展特别兴趣小组。第178–185页（2006） 59. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., 等：Google神经机器翻译系统：弥合人工翻译与机器翻译之间的差距。arXiv（2016） 60. Xu, W., Liu, X., Gong, Y.：基于非负矩阵分解的文档聚类。载于：第26届美国计算机协会信息检索研究与发展国际年会论文集。第267–273页（2003） 61. Yang, W., Zhang, H., Lin, J.：BERT在即席文档检索中的简单应用。arXiv预印本 arXiv:1903.10972（2019） 62. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.：XLNet：用于语言理解的广义自回归预训练。《神经信息处理系统进展》**32** （2019）

63. Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., Zhou, X.：用于语言理解的语义感知 BERT。载于：AAAI。第 34 卷，第 9628–9635 页（2020） 64. Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., Buntine, W.：主题建模与深度神经网络的结合：一项综述。arXiv 预印本 arXiv:2103.00498（2021）