利用标签层级结构，将多答案问答任务与辅助指标用于极端多标签文本分类

Li Wang[1], **Ying Wah Teh**[1*], **Mohammed Ali Al-Garadi**[21]马来亚大学计算机科学与信息技术学院[2]范德堡大学医学中心[1]
wva180023@um.edu.my, tehyw@um.edu.my[2]
mohammed.a.al-garadi@vumc.org

arXiv:2303.01064v1 [cs.CL] 2 Mar 2023

摘要

极端多标签文本分类利用标签层级结构将极端标签划分为多个标签组，从而将任务转化为简单的多组多标签分类任务。当前研究将标签编码为固定长度的向量，因此需要为不同标签组建立多个分类器。问题在于，如何在不牺牲层级结构中标签关系的前提下，仅构建一个分类器。本文将多答案问答任务应用于极端多标签分类，并提出一种辅助分类评估指标。本研究将所提出的方法和评估指标应用于法律领域，并讨论了法律 BERT 的使用以及任务分布。实验结果表明，所提出的层级结构和多答案问答任务能够对 EURLEX 数据集进行极端多标签分类。此外，在多标签分类任务的微调中，领域适配的 BERT 模型在本实验中未显示出明显优势。从理论上说，该方法也适用于零样本学习。

1 引言

极端多标签文本分类（XMTC）是大数据挖掘中的一个持续性问题。它从数以万计的标签集合中为一段文本选择标签。该任务的困难在于极其严重的数据稀疏性、少样本或零样本学习、难以学习标签之间的依赖模式，以及训练和测试的高计算成本 [1]。XMTC 所采用的方法包括一对多（OVA）方法、基于嵌入的方法和划分方法。XMTC 利用标签层级将大量标签划分为多个标签组，从而将极端多标签分类任务转化为简单的多组多标签分类任务。

当前的多标签分类研究将标签编码为一个向量（Chalkidis 等人，2019）。标签向量忽略了标签的语义信息。如果每个标签组中的标签数量不同，就需要为每个标签组建立一个分类器。当标签组数量较多时，就需要构建数量极其庞大的分类器。如何只构建一个分类器，以完成多组多标签分类任务？如果强行将极端标签均匀分组，层级结构中标签之间的关系信息就会丢失。如果强行裁剪标签以平衡标签组的大小，一些标签信息就会丢失。

因此，本文采用多问题回答任务来处理多标签分类。该方法只需构建一个分类器，即可训练标签集长度不同的多标签分类任务。该方法不会牺牲标签信息和标签间关系。此外，该方法在理论上适用于零样本学习。此外，本文还提出了一种用于多答案提问任务分类的辅助评估指标。

根据自然法律语言处理（NLLP）社区 [2]，的观点，法律文本的自动分析很有前景，因为在现实生活中，立法文件篇幅较长、数量庞大，可能使用不同语言，并且带有极端多标签。此外，还有许多具有法律意义的非正式在线文本。因此，NLLP 建议将自然语言处理应用于法律领域，并以法律文本分析为目标。本研究采用所提出的方法和评估指标，以解决极端多标签分类问题中的

法律领域。

*通讯作者

[1] https://analyticsindiamag.com/what-is-extreme-multilabel-text-classification/

[2] https://nllpw.org/

基于 Transformer 的预训练语言模型通过特征工程与分类相结合，自动应用于文本分类。（Minaee 等人，2021年）指出，使用预训练模型包括五个步骤：选择合适的预训练模型、进行领域适配、设计任务专用模型、进行任务专用微调，以及进行模型压缩。该研究还考察了法律领域中基于 BERT 的预训练模型在微调过程中的任务分布。（模型及结果已分享于 [3])

1.1 贡献

本文的贡献如下：

- 本文采用多答案问答任务进行多标签分类。该方法只需构建一个分类器，即可训练标签集长度不同的多标签分类。该方法不会牺牲标签信息和标签间关系。从理论上讲，该方法适用于零样本学习。

- 此外，本文还提出了一种用于多答案问答任务中分类的辅助评估指标。

- 本文将提出的方法和评估指标应用于法律领域的极端多标签分类。本文研究了基于 BERT 的法律领域预训练模型在多标签分类任务中的应用，并研究了使用预训练模型时的任务分布。

2 相关工作

基于 Transformer 的预训练模型（如 BERT）在许多自然语言处理任务中取得了最先进的结果。本文选择 BERT 作为适用于多标签分类的预训练模型。Huggingface 为开发者提供了共享模型的平台，其中包含经过领域适配的模型。法律领域中有多种基于 BERT 的预训练模型。著名的模型包括（Chalkidis 等人，2020年）提出的 Legal BERT，以及（Zheng 等人，2021年）提出的模型。在法律领域的预训练适配方面，（Chalkidis 等人，2020年）试图证明领域预训练的有效性，而（Zheng 等人，2021年）则试图回答领域预训练何时能发挥作用。实验表明，从头开始进行领域预训练和持续预训练都具有竞争力

从头开始和持续预训练工作比较 аҳәыҷ - 相比 BERT 更具优势。然而，在简单任务上进行领域预训练的效果与 BERT 相近。而且，领域预训练对困难任务的帮助大于对简单任务的帮助。然而，当前的法律领域 BERT 模型从领域迁移的角度验证了领域预训练的有效性，但未考虑 - 与任务迁移相关的下游任务分布。本文在多标签分类任务中采用了（ - Chalkidis 等人Chalkidis 等人，2020年）和（Zheng 等人，2021年）提出的法律领域 BERT，以研究任务分布，同时考虑负迁移。

直接使用传统的 MultiLabelBinarizer [4] 和扁平分类（（Silla 和 Freitas，2011）会将极大规模的标签编码为极长的向量。计算机需要大量 RAM 来处理长向量。此外，标签具有语义。MultiLabelBinarizer 编码会抹去这些标签本身的语义。而且，不同的输入编码器和标签编码器会将输入和标签映射到不同的潜在空间。为了获得输入与类别之间的相关性，模型最好将它们嵌入到同一个潜在空间中 [5]。

参照 GLUE [6] 中的 MRPC 数据集，该数据集用于比较句子 1 和句子 2 是否等价，多标签分类任务可以转化为输入与标签之间的等价性／相似度任务，并将标签和输入嵌入同一个潜在空间。显然，将多标签分类转化为基于相似性的二分类是可行的。但由于每个输入都要与每个标签配对，数据集规模会大幅扩大。在极端多标签文本分类中，训练数据集可能涉及数千个标签。将多标签分类转化为二元分类后，训练数据集的规模将是原始数据集大小乘以标签数量，训练将耗时过长。抽取式问答任务 [7] 旨在根据问题检测文本中答案跨度的起始索引和结束索引。通过预测起始索引和结束索引来确定答案的一个优点是

[3] https://github.com/stuwangli/QA-Multilabel-EURLEX

[4] sklearn.preprocessing.MultiLabelBinarizer\T1\ t extemdashscikit-learn1.0.2 文档

[5] https://joeddav.github.io/blog/2020/ 05/29/ZSL.html

[6] https://huggingface.co/datasets/glue/viewer/mrpc/test

[7] https://huggingface.co/docs/transformers/tasks/question_answering

最终答案是文本中从起始索引到结束索引之间的跨度。因此，跨度长度可以根据索引预测动态变化。这适用于多词知识抽取任务。多标签分类任务包含长度可变的多词标签，因此可以利用跨度抽取。

然而，抽取式问答任务只能从文本中预测一个答案跨度。但对于 XMTC，每篇文本可能具有多个标签。如何让抽取式模型进行多跨度抽取/多标签分类？一种处理 SQuAD 数据集中多答案的方法是：在训练数据集中筛除具有多个答案的记录，并在评估时将预测答案与所有答案进行比较，以选择得分最高的答案。[8]

SQuAD 数据集中的多答案记录较少。但在大多数 XMTC 中，文档拥有两个以上的概念标签。因此，实验不能直接采用筛除方法，而需要一种更合适的方式来进行多跨度抽取。

多跨度抽取问题类似于实体抽取问题。因此，根据 Huggingface 中关于实体抽取任务的教程[9]，实验也考虑采用词元级分类。词元级分类任务类似于序列标注任务。在多答案问答中应用序列标注时，输出是一个标签列表，用于表示某个词元是否为正确的类别词元。如果类别列表中存在多个概念，则会有更多词元被标注为类别。在问答任务中，问题不需要标签。（Segal 等人，2019）也将多答案问题视为序列标注问题，并且只对答案进行标签标注。Huggingface [10] 提到，如果一个词元的标签为 -100，则不会将其纳入熵损失计算。因此，实验使用 -100 标注句子1，并使用在多答案问答中表现优于 BIO 的 IO 标注（Segal 等人，2019）来标注句子2。

让我们打开黑盒，看看 BERT-base 模型内部的层。BERT-base 模型包含 BERT 嵌入和 BERT 编码器。BERT 编码器有 12 层范围从第0层到第11层。在 BERT 编码器之后是一个丢弃层，分类器层位于丢弃层之后。BERT 在这些层中包含数百万个参数。（Lee 等人，2019）指出，在基于 Transformer 的模型中，少数几层，尤其是靠近下游任务的底层，在微调过程中受到的影响最大。而顶层提取语言句法，更常用于表示输入。（Lee 等人，2019）Lee 等人（2019）研究了微调过程中受影响的具体层数。通过对 BERT 的实验，他们得出结论：在 BERT基础版中微调最后3至5层，可以达到与整个模型微调相当的性能。因此，在本实验中，概念分类模型冻结了 BERT 中的前11层和嵌入层，而概念分类模型内部的其他层保持解冻，并在构建的数据集上继续进行微调。其他压缩方法，如蒸馏，将在未来的工作中进行探索。

3 方法论

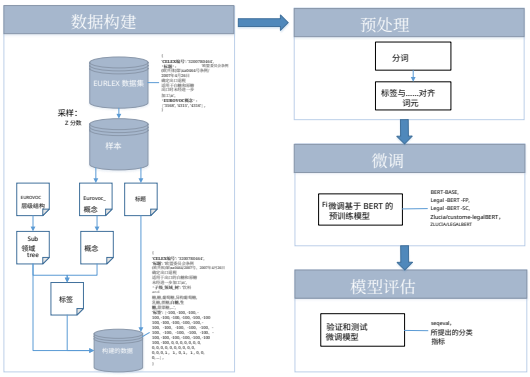

图1：方法论框架

图1展示了该研究方法论的总体框架。第一步是进行数据构建，为多答案问答任务准备数据集。通过对 EURLEX 数据集进行 Z分数采样，获得模拟数据。将带有标签的 EUROVOC 概念映射到 EUROVOC层级结构中的子领域树（/类别列表），以获取多答案的标签。随后，将标签与标题（/文本）结合，作为抽取式多答案问答任务的输入。第二步对输入进行分词，并将标签与词元对齐。预处理数据将在第三步中用于微调基于 BERT 的预训练模型，以完成

[8] https://huggingface.co/course/chapter7/7?fw=pt [9] https://huggingface.co/course/ chapter7/2?fw=pt [10] https://huggingface.co/course/chapter7/2?fw=pt

多答案问答任务。随后，使用 seqeval 以及所提出的辅助分类指标，在验证集和测试样本上评估微调模型的性能。各小节介绍了该方法论中的重要组成部分。数据构建的详细信息将在下一节中说明。

3.1 基于标签层级结构的标签划分

根据（Silla 和 Freitas，2011），层次分类包括构建层级结构（例如创建新的元类），以及依据已构建的层级结构进行分类。构建好层级结构后，分类问题便转变为结构探索问题。

实验使用 EURLEX 数据集进行极端多标签分类。这些极大规模的标签是 EUROVOC 中的概念。EUROVOC [11] 采用层次树结构。除地理领域外，树中的节点之间均为继承关系。目标分类标签，即这些概念，都是树中的叶节点。因此，EUROVOC 树几乎就是“IS-A”层级结构，适合直接用于层次分类。地理领域是一个例外，其中包含名称相同的节点，但这只会影响地理领域的层次分类。我们将在后续需求中考虑这一问题。

3.2 多标签文本分类模型

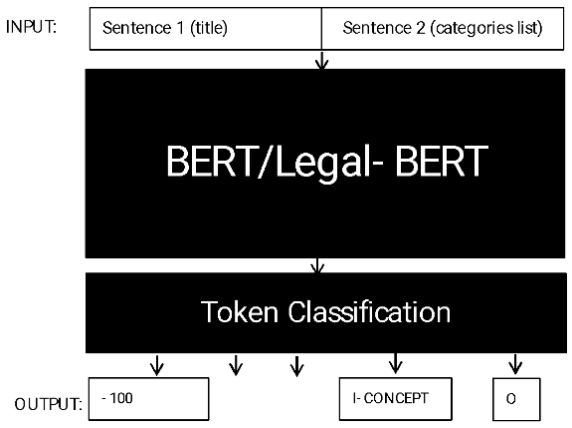

图 2：多标签文本分类模型

多标签文本分类可转化为抽取式多答案问答问题，并可通过词元分类任务加以处理。

[11] https://eur-lex.europa.eu/browse/ eurovoc.html

因此，如图2所示，多标签分类模型直接在基于BERT的预训练模型后堆叠一个词元分类器。模型的输入是将标题（/文本）（句子1）与类别列表（/子领域树）（句子2）填充后组成的一个输入序列。随后，输入序列进入一个基于预训练BERT的模型，以生成其隐藏表示。该隐藏表示将作为下游词元分类器的输入，从输入中的每个词元中选择 [‘I-CONCEPT’、‘O’] 标签。

表 1 展示了用于多标签文本分类的预训练模型。众所周知，基于 Transformer 的预训练模型（例如 BERT）只需经过微调步骤，便能在自然语言处理任务中取得最先进的结果。并且，建议在微调之前对预训练模型进行领域适配。表 1 展示了本方法论中所用预训练模型的详细信息。如表所示，BERT（Devlin 等人，2018年）使用 Book Corpus 和维基百科等通用知识进行预训练。Legal-BERT-SC 和 Legal-BERT-FP 是文献（Chalkidis 等人，2020年）提出的法律领域专用模型。Legal-BERT-SC 使用 12 GB 英文法律文本从头开始预训练，具体包括 EUR-Lex、ECHR 和 EDGAR。Legal-BERT-FP 则以 BERT 为初始模型，使用相同的法律文档进行持续预训练。zlucia/custom-legalbert 和 zlucia/legalbert 是文献（Zheng 等人，2021年）提出的法律领域 BERT 模型。zlucia/custom-legalbert 使用 CaseHOLD 数据集，并结合领域专用法律词汇从头开始预训练；zlucia/legalbert 则使用相同的数据集进行持续预训练，不添加额外的法律词汇。通过领域知识进行持续预训练的法律领域 BERT 模型，均以 bert-base-uncased 模型为初始模型。这些模型已发布在 huggingface.co 和 Archive.org 上。本研究选用 Legal-BERT-FP/bert-base-500k 模型作为 Legal-BERT-FP。法律领域 BERT 模型通过从头开始训练或持续预训练，执行掩码语言建模（MLM）这一自监督学习任务。本文中，BERT 和 BERT-BASE 均为 BERT-base-uncased 的简称。zlucia/custom-legalbert 也称为 ZLUCIA/CUSTOM-LEGALBERT 和 Zlucia/custome-legalBERT。同样，zlucia/legalbert 也称为 ZLUCIA/LEGALBERT 和 Zlucia/legalBERT。

表2比较了来自（Chalkidis

模型	数据集	预训练方法	初始化模型	自监督任务	发布链接。
BERT-BASE-UNCASED (Devlin 等人，2018年)	BookCorpus、英文维基百科。	从头开始	-	掩码语言模型、下一句预测	https://huggingface.co/bert-base-uncased
LEGAL-BERT-FP (Chalkidis 等人，2020年)	EUR-Lex、ECHR、EDGAR，12 G	持续预训练	bert-base-uncased	MLM	https://archive.org/details/legal_bert_fp
LEGAL-BERT-SC (Chalkidis 等人，2020年)	EUR-Lex、ECHR、EDGAR，12 G	从头开始	-	MLM	https://huggingface.co/nlpaueb/legal-bert-base-uncased
ZLUCIA/CUSTOM-LEGALBERT（Zheng 等人，2021年）	CaseHOLD 数据集	从头开始 + 领域专用法律词汇	-	MLM	https://huggingface.co/zlucia/custom-legalbert
ZLUCIA/LEGALBERT(Zheng 等人，2021年)	CaseHOLD 数据集	持续预训练	bert-base-uncased	MLM	https://huggingface.co/zlucia/legalbert

表 1：法律领域中基于 BERT 的预训练模型

模型	优势	劣势
LEGAL-BERT-FP	通过持续预训练适应领域。 网格搜索会影响性能。 性能优于 BERT，尤其是在低资源测试环境中。 在 ECHR 案件和美国合同方面适应得更快、更好。	未考虑任务分布。
LEGAL-BERT-SC	通过从头预训练来适应领域。 网格搜索会影响性能。 表现优于 BERT，尤其是在低资源测试环境中。 在 EURLEX57K、ECHR-CASES 和 CONTRACTS-NER 上的困惑度（PPT）较低。	未考虑任务分布。 与持续预训练模型相比，未考虑其在其他领域任务中的泛化能力。
ZLUCIA/CUSTOM-LEGALBERT	通过持续预训练适应领域。 对于领域特异性不强的中等难度任务，领域预训练可以有所帮助 对于难度极高且具有领域特异性的任务，收益最为显著。 网格搜索会影响性能。	不考虑任务分布，只考虑问答任务中的提示长度。
ZLUCIA/LEGALBERT	采用具有特定法律词汇的领域，并从头预训练。 领域预训练有助于处理中等难度和困难任务。 在推翻先例、服务条款和 CaseHOLD 任务上取得了优于 Legal BERT 的测试性能。	在问答任务中未考虑任务分布，仅考虑提示规模。 与持续预训练模型相比，未考虑在其他领域任务中的泛化能力。 网格搜索会影响性能。

表2：法律领域中基于 BERT 的预训练模型对比表

等人，2020年）和（Zheng 等人，2021年）的法律领域 BERT 的优势和不足。从头预训练和持续预训练的效果比 BERT 更具竞争力。然而，在简单任务上进行领域预训练时，其表现与 BERT 相似。此外，领域预训练对复杂任务的帮助大于对简单任务的帮助。然而，当前的法律领域 BERT 模型从领域迁移的角度测试了领域预训练的有效性，却未考虑与任务迁移相关的下游任务微调。本文将研究法律领域 BERT 在具有不同下游任务结构的多标签分类任务微调中的有效性。除了所提出的多答案问答下游任务外，本文还研究了序列分类和相似度任务。由于多标签分类这一概念只能通过多答案问答任务结构来设计，本文加入了领域多标签分类任务，该任务具有固定的领域类别列表，用于研究预训练模型与不同下游任务在微调过程中的关系。

3.3 评估指标

Huggingface [12] 指出，令牌分类的评估使用 seqeval [13]，这是一个用于序列标注的 Python 框架。评估的默认模式适用于 IO 标签。它支持精确率、召回率、F 分数和准确率等指标。由于 seqeval 尝试处理分块任务，因此会严格识别块的边界。精确率、召回率和 F 分数均基于对完全原始块的映射进行计算。

然而，抽取式多标签分类任务更注重识别正确的类别，而非边界检测。如果只抽取类别标签中的部分词元，由于模型关注的是正确类别，因此在该任务上的表现仍然不错。因此，抽取式多标签文本分类的评估更倾向于判断是否从正确类别中抽取信息（例如词元），而非是否抽取了完整准确的类别片段。为此，研究提出了另一种知识抽取评估方法：在应用 seqeval 之前，将类别边界添加到预测的类别词元中。相似度任务使用 GLUE/MRPC 指标 [14] 进行评估。

4 数据集

Huggingface 数据集提供了 EUR-Lex 的预览。该数据集包含四列，分别是文档 ID、标题、文档文本和概念标签。标题作为第一段包含在文档文本中。每份文档都有多个概念标签，这些标签根据 EUROVOC 列表标注，仅使用概念标签 ID，而非概念短语。大多数标注的概念都出现在标题中，或以同义词的形式出现在标题中。EUR-Lex 中最长的文档超过 20,000 个词元，不过大多数文档不到 5,000 个词元。少数文档约有 512 个词元，这一长度适合作为 BERT 输入。因此，

[12] https://huggingface.co/course/ 第7章/2

[13] https://github.com/chakki-works/seqeval

[14] https://huggingface.co/course/chapter3/4?fw=pt

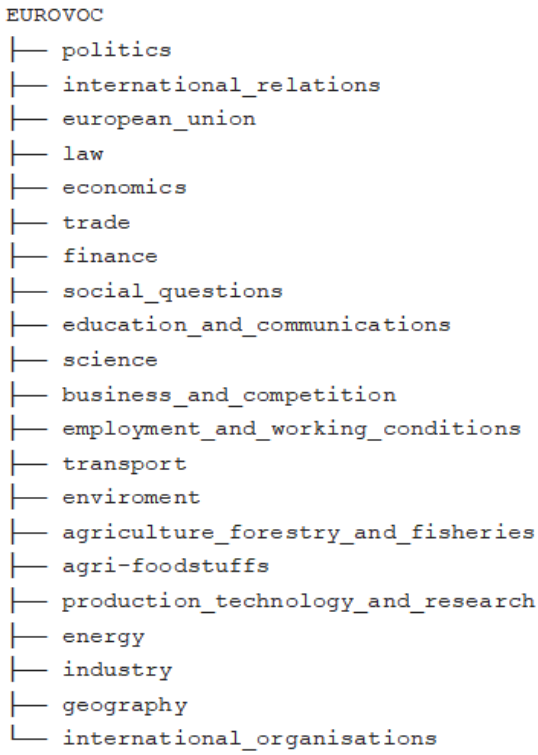

图3：EUROVOC中的领域

本实验使用标题作为多标签文本分类的文本。数据集卡片提到，EUR-Lex 包含 45,000 份训练文档、6,000 份验证文档和 6,000 份测试文档。

EURLEX 数据集中有数千个概念类别。如果将所有类别组成一个类别列表，作为输入中的第二个句子，BERT就会遇到输入过长的问题。因此，本实验利用EUROVOC层级结构将概念类别划分为若干组。

由于网页内容动态变化，EUROVOC树是从EUROVOC网站手动下载的。该树存储在本地，并使用anytree软件包构建 [15]。构建树时，添加了一个名为“eurovoc”的根节点，作为整棵树的祖先节点。为便于将节点名称映射到概念标签，所有节点名称均以小写形式存储。包含根节点在内，树高为8。第一级为“eurovoc”，第二级为EUROVOC中的21个领域，如图3所示。

由于 EUROVOC 采用树结构，节点分别属于不同的子树；后代节点与父节点之间存在内部继承关系（“IS-A”关系）或组合关系（“HAS-A”关系）

[15]https://anytree.readthedocs.io/en/ latest/

{
'celex_id': '32007R0464',
'title': 'Commission Regulation (EC) No\xa0464/2007 of 26 April 2007 fixing the export refunds on white and raw sugar exported without further processing\n',
'eurovoc_concepts': ['3568', '4315', '4316'],
'concepts': 'export refund', 'white sugar', 'raw sugar',
}

图4：带有‘celex_id’的‘32007R0464’数据记录

。每个子树都类似一个领域簇。子树从父节点开始，以直接或间接连接到该父节点的叶节点为终点。从叶节点到父节点的路径上若存在其他节点，这些节点也属于该子树。层级结构中的每个节点（叶节点除外）都可以作为子树的父节点。然而，如果子树过大，包含的节点过多，类别列表就会过长，难以处理；如果子树过小，则模型需要大量子树才能覆盖所有概念标签。因此，选择子树时应兼顾这两种情况。

因此，表3展示了在 EUROVOC 层级结构中选择不同层级作为父节点时，对子树数量以及子树中节点数量的分析。综合考虑这两种情况，研究选取第3层的节点作为子树的父节点。EUROVOC 指出，EUROVOC 中的概念分布在21个领域和127个子领域中。第3层的父节点即这127个子领域。因此，研究将这些子领域对应的子树称为子领域树。

为微调概念分类模型，EURLEX 数据集利用子领域树构建类别列表，并在序列标注中使用 [-100、I-CONCEPT、O] 标记概念。图1展示了从 EURLEX 数据集中选取的两列，具体而言，是标题和概念。随后，在每棵子领域树中遍历概念，并为每个标题与类别列表配对生成序列标签。考虑到处理大型构建数据集所需的计算时间，实验通过筛除空答案样本，将该任务视为不包含空答案样本的问答任务。换言之，如果标题对应的概念不在某些子领域树中，就会从构建的数据集中删除该标题与类别列表的配对。下面给出一个示例来说明这一过程。

图4展示了 EURLEX 数据集中的记录。该记录的“CELEX 编号”为‘32007R0464’。并且_

父节点层级结构	子树数量	子树中的平均节点数	最大节点数	最小节点数
1	1	8274	-	-
2	21	394	1597	137
3	127	65	532	1
4	547	15	80	1

表3：EUROVOC层级结构分析

```
Node('/eurovoc/agri-foodstuffs/beverages and sugar')
beverages and sugar
├── sugar
│   ├── glucose
│   ├── isoglucose
│   ├── lactose
│   ├── sucrose
│   ├── white sugar
│   ├── raw sugar
│   ├── beet sugar
│   └── cane sugar
└── beverage
    ├── alcoholic beverage
    │   ├── alcohol
    │   ├── liqueur
    │   ├── aperitif
    │   ├── wine
    │   │   ├── flavoured wine
    │   │   ├── white wine
    │   │   ├── local wine
    │   │   ├── wine of superior quality
    │   │   ├── table wine
    │   │   ├── bottled wine
    │   │   ├── fortified wine
    │   │   ├── sparkling wine
    │   │   ├── ros wine
    │   │   ├── red wine
    │   │   ├── still wine
    │   │   └── champagne
    │   ├── beer
    │   ├── cider
    │   └── spirits
    └── non-alcoholic beverage
        ├── fruit juice
        ├── vegetable juice
        ├── aerated drink
        └── mineral water
```

图5：父节点为“饮料和糖”的子领域树

“eurovoc_concepts”是“eurovoc_concepts.jsonl”文件中的概念“ID”。为了获取 eurovoc_concepts 的名称/标题，实验首先将每个“ID”映射到 jsonl 文件中对应的“标题”。因此，该数据记录包含 3 个概念 ID，映射后会得到 3 个不同的概念。

然后，将在 EUROVOC 的子领域树中遍历每个概念。例如，在图 5 中，“白糖”和“原糖”位于以“饮料和糖”为父节点的子领域树上。


```
{
'celex_id': '32007R0464',
'title': 'Commission Regulation (EC) No\xa0464/2007 of 26 April
2007 fixing the export refunds on white and raw sugar exported
without further processing\n',
'sub_domainstree': 'beverages and
sugar,sugar,glucose,isoglucose,lactose,sucrose,white sugar,raw
sugar,beet sugar,cane sugar,beverage,alcoholic
beverage,alcohol,liqueur,aperitif,wine,flavoured wine,white
wine,local wine,wine of superior quality,table wine,bottled
wine,fortified wine,sparkling wine,ros wine,red wine,still
wine,champagne,beer,cider,spirits,non-alcoholic beverage,fruit
juice,vegetable juice,aerated drink,mineral water,',
'concepts': 'white sugar,raw sugar',
}
```


图6：‘32007R0464’标题与‘饮料和糖’子领域树的配对


```
{
'celex_id': '32007R0464',
'labels': [-100, -100, -100, -100, -100, -100, -100, -100, -
100, -100, -100, -100, -100, -100, -100, -100, -100, -100, -
100, -100, -100, -100, -100, -100, -100, -100, -100, -100, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, -100, -100, -100,…],
}
```


图7：输入内容的标签；输入由‘32007R0464’标题和‘饮料和糖’子领域树类别列表组成

图 6 展示了子领域树的类别列表，以及“白糖”和“原糖”在类别列表中的位置。在类别列表中，各类别以逗号分隔，便于后续进行序列标注。在原文中插入标点符号这一数据扩充方法也在（Karimi 等人，2021年）中提及。

图7展示了输入内容对应的序列标注标签。该输入由‘32007R0464’标题和‘饮料和糖’子领域树类别列表组成。标题标注为 -100，类别列表中的概念标注为 1，其他内容标注为 0。类别列表之后还有一些 -100，它们是填充词元的标签。

表4表示，数据构建后，训练数据集从45000扩大到205053。由于数据量极大，实验从数据集中随机选择样本，并在

	训练	验证	Test
数据集	45000	6000	6000
构建后的数据集	205053	27298	27149
随机样本	381	362	362
构建后的样本	1708	1650	1648

表4：数据集大小

随机样本上进行实验。样本选择在数据构建步骤之前进行。样本量通过基于Z分数的在线样本量计算器 [16] 计算得出。置信水平设为95%，误差范围设为5%。根据计算结果，训练样本量为45000中的381个。验证样本量和测试样本量均为6000中的362个。数据构建后，训练、验证和测试数据样本分别为1708、1650和1648个。随机样本选择参考了Hugging Face [17] 和Pandas [18]。

多答案问答任务首先使用负样本进行训练。训练一个轮次需要超过20小时，但模型在验证中的表现非常差。考虑到计算时间成本，本实验将该任务作为不含空答案样本的问答任务处理：过滤掉空答案样本，并对数据集进行6轮训练。如何理解和利用负样本仍是一个问题，将在未来进行研究。

5 评估结果与讨论

表5展示了概念多标签分类的验证和测试结果。由于概念多标签分类中的标签组长度各不相同，下游任务采用了方法论中提出的多答案问答任务，以两个句子作为输入，并使用令牌分类对句子2进行类别标注。评估指标包括训练损失、验证损失以及总体指标——精确率、召回率、F1分数和seqeval中的准确率。

表5中五种预训练模型的训练损失和验证损失均低于10%。这表明，采用冻结层的 BERT 模型适用于多答案问答下游任务，能够适应训练集和验证集的分布。多标签分类任务十分复杂，输入中的句子1和句子2都是变量，而且该任务要求预测句子2中每个词元的标签。因此，需要足够多的参数来模拟任务分布。BERT 和基于 Transformer 的模型包含数百万甚至更多参数，这与规模较小的浅层模型不同。大型模型能够表示复杂的任务分布。

验证和测试结果表明，在五个预训练模型中，Zlucia/custome-legalBERT 无法通过微调完成该任务，而其他四个模型的表现从好到差依次为 Legal-BERT-FP、BERT-base-uncased 和 Zlucia/legalBERT。在第 6 个训练轮次时，Legal-BERT-FP 的 F 分数比 BERT-base-uncased 高约 5%，精确率高 6%，召回率高 4%。然而，Zlucia/legalBERT 同样是经过领域适配并持续预训练的预训练模型，但微调后其 F 分数却低于 BERT-base-uncased。Legal-BERT-FP 与 Zlucia/legalBERT 的区别在于，Zlucia/legalBERT 使用带有掩码语言模型任务的 CaseHOLD 数据集进行训练，而 Legal-BERT-FP 使用带有掩码语言模型任务的 EUR-Lex、ECHR 和 EDGAR 数据集进行训练。

Legal-BERT-FP 和 Zlucia/legalBERT 的表现优于 Zlucia/custome-legalBERT 和 Legal BERT-SC，这表明持续预训练优于从头训练。从头开始进行预训练会丢弃 BERT-base-uncased 的预训练知识，而利用 BERT-base-uncased 的预训练知识则能够提升性能。

尽管 Legal-BERT-FP 的表现优于 BERT，但两者的 F 分数相近。这表明，经过精细或少量微调，BERT-base-uncased 可以成为表现出色的少样本学习器，达到与领域持续预训练 BERT 模型相近的性能。

BERT 和 Legal-BERT-FP 的 F 分数均维持在 30% 左右，精确率低于 50%。分析 Legal-BERT-FP 在第 6 个训练轮次提取的概念后发现，存在许多空提取。这可能是因为训练样本量太小，无法提供足够的信息供模型学习。而且，大多数提取的概念只包含类别中的部分词元。

[16] https://www.qualtrics.com/experience-management/research/determine-sample-size/ [17] https://huggingface.co/course/ chapter5/3?fw=pt [18] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample. html

				验证				测试			
模型	训练轮次	训练损失	验证损失	精确率	召回率	F 分数	准确率	精确率	召回率	F 分数	准确率
BERT-BASE-UNCASED	1	8.03	4.65	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	4.38	3.86	15.65	5.96	8.63	99.03	15.01	5.05	7.55	99.02
	3	3.92	3.58	30.59	15.85	20.88	99.10	29.62	14.07	19.08	99.08
	4	3.67	3.47	32.53	16.29	21.71	99.12	31.66	14.65	20.03	99.09
	5	3.54	3.40	35.25	20.55	25.96	99.13	34.14	18.58	24.07	99.11
	6	3.44	3.39	35.38	23.22	28.04	99.12	34.59	21.79	26.73	99.11
LEGAL-BERT-FP	1	8.81	4.76	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	4.44	3.87	21.29	3.83	6.49	99.03	17.38	2.96	5.06	99.02
	3	3.79	3.54	31.73	18.18	23.11	99.07	30.45	16.89	21.72	99.06
	4	3.49	3.33	39.18	17.55	24.24	99.14	39.62	17.13	23.92	99.12
	5	3.33	3.27	41.93	24.67	31.07	99.15	40.42	22.51	28.92	99.13
	6	3.22	3.27	41.16	27.29	32.82	99.15	39.30	25.13	30.66	99.13
LEGAL-BERT-SC	1	7.40	5.06	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	5.01	4.84	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	3	4.84	4.66	43.33	0.63	1.24	99.02	35.71	0.73	1.43	99.01
	4	4.68	4.48	35.51	2.38	4.45	99.03	40.60	2.62	4.92	99.02
	5	4.55	4.39	27.44	3.54	6.27	99.02	32.18	4.08	7.24	99.02
	6	4.46	4.37	26.87	4.36	7.51	99.03	30.15	4.76	8.21	99.02
ZLUCIA/CUSTOM-LEGALBERT	1	6.72	5.10	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	5.17	4.97	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	3	5.07	4.87	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	4	4.97	4.80	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	5	4.94	4.76	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	6	4.90	4.74	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
ZLUCIA/LEGALBERT	1	7.83	4.91	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	4.68	4.18	0.00	0.00	0.00	99.00	0.00	0.00	0.00	98.99
	3	4.15	3.87	14.08	3.78	5.96	99.03	12.21	2.77	4.51	99.02
	4	3.89	3.69	23.74	7.56	11.47	99.07	20.18	5.58	8.74	99.05
	5	3.71	3.60	28.57	13.96	18.76	99.07	25.85	11.40	15.82	99.06
	6	3.64	3.57	29.52	14.15	19.14	99.08	26.59	11.55	16.10	99.07

表5：使用 seqeval 进行概念多标签分类多答案问答下游任务的验证和测试结果（%）

由于 seqeval 指标更关注边界检测，因此精确率、召回率和 F 分数都较低。

从第一个训练轮次开始，准确率就很高。然而，在第一个训练轮次中，精确率、召回率和 F 分数均为零。其原因可能是样本数量较少，且标签中‘I-CONCEPT’所占比例很小。因此，如果模型将所有标签都预测为‘O’，并且由于标签内部‘O’所占比例很高，准确率仍然会很高。该准确率可能表明，序列标注中的‘O’标签约占99%。

结果表明，所提出的多答案问答模型满足概念多标签分类任务的建模要求。为了获得更好的性能，该模型在微调时需要更多数据集。

从人的角度来看，人们做多选题时，有时只需从答案的一部分而非整个答案片段中选出正确选项。由于这项研究关注的是正确类别预测，而非跨度边界检测，因此即使模型只识别出正确类别中的一个词元，在这项任务中仍然表现良好。因此，评估应判断模型是否从正确类别中挖掘出信息或词元。

为了满足评估需求，研究提出了第二种验证指标——分类指标：在应用 seqeval 之前，根据预测的词元补全类别边界。由于类别列表中的各个类别以“,”分隔，因此很容易根据预测的词元定位完整的类别标签。这与人们做多选题时的情况类似，不同选项之间会用空格或符号分隔。

表6展示了所提出的分类指标评估结果。总体精确率、召回率和 F 分数均优于 seqeval 验证结果。最佳精确率超过60%，F 分数接近50%。在原始文本中插入标点符号不仅有助于数据构建，也有助于分类任务评估。

验证和测试结果表明，Zlucia/custome-legalBERT 无法通过微调样本模拟该任务；其余四种模型的表现从好到差依次为 Legal-BERT-FP、BERT-BASE 和 Zlucia/legalBERT。此外，Legal-BERT-FP 和 BERT-base-uncased 在验证集和测试数据集上均取得了有竞争力的 F 分数。在训练轮次 5 时，Legal-BERT-FP 在验证集上的表现略优于 BERT，在测试集上的表现则略逊于 BERT。

抽取出的答案并非独立于类别在类别列表中的位置。

		验证				测试			
模型	**训练轮次**	**精确率**	**召回率**	**F 分数**	**准确率**	**精确率**	**召回率**	**F 分数**	**准确率**
BERT-BASE-UNCASED	1	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	53.10	19.53	28.56	99.01	54.26	17.61	26.59	99.01
	3	61.81	31.46	41.70	99.07	63.72	29.74	40.56	99.08
	4	66.31	32.62	43.73	99.12	67.77	30.81	42.36	99.11
	5	62.86	36.02	45.79	99.09	65.21	34.84	45.41	99.10
	6	60.00	38.68	47.04	99.06	60.79	37.46	46.35	99.07
LEGAL-BERT-FP	1	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	58.15	10.37	17.61	99.03	60.76	10.14	17.38	99.03
	3	53.83	30.34	38.81	99.02	52.90	28.77	37.27	99.00
	4	68.54	30.10	41.83	99.14	68.03	29.11	40.77	99.12
	5	63.25	36.45	46.25	99.11	61.80	33.67	43.59	99.08
	6	61.60	40.04	48.53	99.12	59.69	37.36	45.96	99.07
LEGAL-BERT-SC	1	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	3	96.55	1.36	2.68	99.02	90.00	1.75	3.43	99.01
	4	63.64	4.07	7.65	99.01	68.25	4.17	7.86	99.01
	5	51.38	6.30	11.23	98.99	58.94	7.04	12.57	99.00
	6	52.94	8.29	14.33	99.00	55.91	8.49	14.74	98.99
ZLUCIA/CUSTOM-LEGALBERT	1	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	3	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	4	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	5	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	6	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
ZLUCIA/LEGALBERT	1	0.00	0.00	0.00	99.01	0.00	0.00	0.00	98.99
	2	0.00	0.00	0.00	99.00	0.00	0.00	0.00	98.99
	3	54.03	13.96	22.19	99.03	57.05	12.57	20.60	99.03
	4	61.35	19.00	29.02	99.07	64.56	17.32	27.31	99.07
	5	55.61	26.66	36.04	99.02	56.26	23.97	33.62	99.02
	6	58.38	27.53	37.42	99.05	60.48	25.47	35.85	99.06

表6：采用所提出的分类指标进行多答案问答下游任务中的概念多标签分类，其验证和测试结果（%）

如果同一标题和同一领域树中的类别在类别列表中随机重新排序，抽取出的答案可能会有所不同。然而，在分类中，类别的顺序并不重要。例如，验证数据集中 CELEX 编号为“31999L0010”的记录，抽取出了“食品”，但_在类别列表重新排序后，得到的是“食品”和“加工食品”。(Chalkidis 等人，2019）Ko等（2020）也提到了抽取式问答中的位置偏差。所提出的多标签分类任务还有一个问题，即输入过长，因为 BERT 的输入长度有限。

表7展示了采用方法论中提出的多答案问答任务作为下游任务时，领域多标签分类的验证和测试结果。领域多标签分类具有固定的领域标签组，而概念多标签分类的标签组长度则各不相同。评估指标包括训练损失、验证损失，以及总体指标——精确率、召回率、F1分数和 seqeval 中的准确率。表7展示了微调后的领域分类模型的验证结果，具体包括 BERT-BASE 和四个法律领域 BERT。

表7中的训练损失和验证损失处于五种基于 BERT 的预训练模型的 20% 至 40% 之间，这些模型采用冻结层并用于多答案问答下游任务。领域多标签分类任务仅将句子1作为输入变量，而句子2为常量。因此，该任务的复杂度低于概念多标签分类任务。

验证和测试结果表明，五种预训练模型均能模拟领域多标签分类任务，其表现从好到差依次为 BERT-base-uncased、Zlucia/legalBERT、Zlucia/custome-legalBERT、Legal-BERT-FP 和 Legal BERT-SC。持续预训练优于从头训练。BERT-BASE 表现最佳，Zlucia/legalBERT 和 Zlucia/custome-legalBERT 的表现优于 Legal-BERT-FP 和 Legal BERT-SC。

表8展示了在采用多答案问答任务作为下游任务的领域多标签分类中，分类指标的结果。验证和测试结果表明，

				验证				测试			
模型	**训练轮次**	**训练损失**	**验证损失**	**精确率**	**召回率**	**F 分数**	**准确率**	**精确率**	**召回率**	**F 分数**	**准确率**
BERT-BASE-UNCASED	1	32.75	26.10	13.70	7.99	10.09	88.75	11.64	6.86	8.63	88.26
	2	24.80	22.71	38.09	30.84	34.08	90.84	34.67	28.62	31.35	90.36
	3	22.39	21.40	50.87	44.77	47.63	91.49	48.85	43.18	45.84	91.19
	4	21.27	20.86	52.60	47.24	49.78	91.65	51.57	45.81	48.52	91.43
	5	20.71	20.46	52.00	46.39	49.03	91.87	50.53	44.71	47.44	91.63
	6	20.21	20.23	52.87	45.45	48.88	91.99	51.32	44.28	47.55	91.76
LEGAL-BERT-FP	1	35.09	27.68	26.52	8.16	12.48	88.02	25.07	7.71	11.79	87.65
	2	26.44	25.18	55.51	33.81	42.03	88.34	52.35	32.09	39.79	87.86
	3	24.78	24.40	43.01	33.98	37.97	88.53	39.67	32.18	35.53	88.16
	4	23.89	23.47	38.88	35.34	37.03	89.44	36.23	33.53	34.83	89.10
	5	23.10	22.52	44.64	32.54	37.64	90.83	40.63	30.82	35.05	90.57
	6	22.53	22.16	43.96	33.98	38.33	90.97	40.99	32.35	36.16	90.84
LEGAL-BERT-SC	1	35.04	27.77	52.76	16.23	24.82	88.12	48.21	14.82	22.67	87.68
	2	26.38	25.41	55.06	33.31	41.50	88.35	52.14	31.92	39.60	87.84
	3	25.03	25.02	38.46	33.98	36.08	88.14	35.86	31.67	33.63	87.61
	4	24.49	24.50	37.85	33.47	35.53	88.62	34.74	31.41	32.99	88.15
	5	24.15	23.90	44.21	31.44	36.74	89.80	42.06	29.81	34.89	89.31
	6	23.77	23.71	43.50	31.86	36.78	90.11	40.02	30.06	34.33	89.58
ZLUCIA/CUSTOM-LEGALBERT	1	36.80	31.07	0.00	0.00	0.00	87.54	0.00	0.00	0.00	87.81
	2	30.37	27.33	5.23	2.04	2.93	88.10	5.66	2.29	3.26	87.75
	3	28.03	26.30	39.41	24.98	30.58	88.50	35.14	23.12	27.89	87.71
	4	27.26	26.06	47.60	32.80	38.83	88.49	44.36	30.65	36.25	87.76
	5	27.13	25.77	48.10	31.27	37.90	88.54	44.75	29.21	35.35	87.80
	6	26.74	25.79	48.28	33.39	39.48	88.47	44.16	31.08	36.48	87.73
ZLUCIA/LEGALBERT	1	33.89	27.43	0.28	0.09	0.13	88.27	0.28	0.08	0.13	87.91
	2	26.31	24.71	49.60	15.89	24.07	88.57	45.29	14.65	22.14	88.21
	3	24.13	22.88	44.00	34.24	38.51	90.32	41.50	32.26	36.30	89.96
	4	22.64	21.78	47.62	40.02	43.49	90.98	45.21	39.12	41.94	90.68
	5	21.77	21.09	47.44	40.10	43.46	91.29	45.34	39.12	42.00	90.99
	6	21.35	20.82	48.36	40.19	43.90	91.47	45.95	38.95	42.16	91.21

表7：使用 seqeval 进行多答案问答下游任务领域多标签分类的验证和测试结果（%）

		验证				测试			
模型	**训练轮次**	**精确率**	**召回率**	**F 分数**	**准确率**	**精确率**	**召回率**	**F 分数**	**准确率**
BERT-BASE-UNCASED	1	53.08	19.80	28.84	88.26	49.66	18.37	26.82	87.83
	2	55.83	39.85	46.50	88.55	52.82	38.10	44.27	88.02
	3	60.10	50.55	54.91	90.31	58.75	49.45	53.70	90.08
	4	60.99	52.34	56.33	90.74	59.68	50.89	54.94	90.63
	5	61.92	52.51	56.83	91.15	60.18	51.31	55.39	90.87
	6	62.55	51.66	56.58	91.37	61.06	50.72	55.41	91.13
LEGAL-BERT-FP	1	49.49	16.40	24.63	87.99	46.37	15.16	22.85	87.58
	2	54.14	33.90	41.69	88.27	50.94	32.09	39.38	87.76
	3	52.87	40.70	45.99	88.14	50.70	39.71	44.54	87.73
	4	52.64	43.25	47.48	88.12	50.83	41.32	45.59	87.76
	5	57.41	38.83	46.33	89.58	55.21	38.10	45.09	89.19
	6	56.70	40.61	47.33	89.60	55.41	39.46	46.09	89.47
LEGAL-BERT-SC	1	50.26	16.48	24.82	88.02	47.37	15.24	23.06	87.62
	2	54.42	33.47	41.45	88.28	51.64	32.01	39.52	87.80
	3	49.62	43.84	46.55	86.97	47.85	42.42	44.97	86.58
	4	51.15	43.33	46.92	87.45	48.97	42.08	45.26	86.99
	5	55.05	34.24	42.22	88.33	52.25	32.43	40.02	87.93
	6	55.50	36.02	43.69	88.52	52.77	34.72	41.88	88.12
ZLUCIA/CUSTOM-LEGALBERT	1	0.00	0.00	0.00	87.52	0.00	0.00	0.00	87.79
	2	50.65	16.65	25.06	87.98	45.72	15.83	23.52	87.23
	3	54.77	32.20	40.56	88.27	49.79	30.57	37.88	87.56
	4	52.49	35.77	42.55	87.92	49.82	34.38	40.68	87.53
	5	54.45	34.32	42.11	88.24	50.07	32.18	39.18	87.59
	6	52.68	35.85	42.67	87.90	49.33	34.55	40.64	87.44
ZLUCIA/LEGALBERT	1	51.74	16.40	24.90	88.06	48.39	15.24	23.18	87.65
	2	51.42	16.91	25.45	88.03	48.08	15.92	23.92	87.68
	3	57.24	41.63	48.20	89.68	54.92	40.64	46.72	89.36
	4	58.40	46.64	51.87	90.31	56.59	46.15	50.84	89.95
	5	60.27	48.60	53.81	90.85	57.97	47.76	52.37	90.49
	6	60.99	48.09	53.78	91.08	58.70	48.01	52.82	90.74

表8：使用所提出的分类指标进行多答案问答任务的领域多标签分类验证和测试结果（%）

				验证		测试	
模型	**训练轮次**	**训练损失**	**验证损失**	**F 分数**	**准确率**	**F 分数**	**准确率**
BERT-BASE-UNCASED	1	57.80	48.56	42.24	7.73	39.25	9.94
	2	44.98	41.11	43.52	8.01	40.79	10.22
	3	39.90	38.05	43.59	8.01	41.59	10.22
	4	37.54	36.52	44.12	8.01	41.14	10.22
	5	36.36	35.78	45.56	8.29	42.93	10.22
	6	35.76	35.57	45.29	8.29	42.41	10.22
LEGAL-BERT-FP	1	61.34	50.68	44.38	7.73	42.69	10.22
	2	46.36	41.50	42.76	8.01	40.45	10.22
	3	40.04	37.88	42.90	8.01	40.53	10.22
	4	37.42	36.34	43.06	8.01	40.53	10.22
	5	36.26	35.63	43.14	8.01	40.60	10.22
	6	35.68	35.41	43.13	8.01	40.49	10.22
LEGAL-BERT-SC	1	60.89	50.65	41.20	5.25	40.11	4.70
	2	46.61	41.33	42.64	7.46	38.79	9.94
	3	40.17	37.95	43.30	8.01	40.29	10.22
	4	37.65	36.50	43.13	8.01	40.33	10.22
	5	36.48	35.80	43.40	8.01	40.54	10.22
	6	35.97	35.58	43.33	8.01	40.51	10.22
ZLUCIA/CUSTOM-LEGALBERT	1	53.09	42.38	43.40	7.46	41.69	10.22
	2	39.67	36.94	42.51	8.01	40.06	10.22
	3	36.26	35.34	42.26	7.73	40.17	10.22
	4	35.09	34.75	42.34	7.73	40.64	10.22
	5	34.75	34.44	42.69	8.01	40.83	10.22
	6	34.44	34.36	42.50	7.73	40.78	10.22
ZLUCIA/LEGALBERT	1	61.10	50.76	43.00	8.01	40.26	10.22
	2	45.79	41.05	42.68	8.01	40.04	10.22
	3	39.54	37.70	43.42	8.01	39.91	9.94
	4	37.21	36.34	43.77	8.01	40.28	10.22
	5	36.17	35.68	43.57	8.01	40.32	10.22
	6	35.60	35.46	43.55	8.01	40.39	10.22

表9：V　领域多标签序列分类的验证与测试结果（%）　　n
下游　　任务

五种预训练模型的表现从好到差依次为 BERT-base-uncased、Zlucia/legalBERT、Legal-BERT-FP、Legal BERT-SC 和 Zlucia/custome-legalBERT。BERT-base-uncased 的表现最佳，且持续预训练优于从头训练。此外，Zlucia/legalBERT 优于 Legal-BERT-FP。Legal BERT-SC 优于 Zlucia/custome-legalBERT。

表9展示了在序列分类下游任务中，领域多标签分类的验证和测试结果。序列分类将恒定领域标签组编码为一个向量，而不利用标签语义信息。评估指标包括训练损失、验证损失、F1分数和准确率。

表9中，采用冻结层和序列分类下游任务的五个基于 BERT 的预训练模型，其训练损失和验证损失介于 30% 至 60% 之间。领域多标签分类任务仅将句子1作为输入，因为输入是一个变量。因此，该任务的复杂度低于多答案问答任务。

验证和测试结果显示，准确率非常低，约为 10%。这五个预训练模型在分配任务方面的表现相近，其中 BERT-BASE-UNCASED 略胜一筹。

表10展示了在下游相似度任务中，领域多标签分类的验证和测试结果。评估指标包括训练损失、验证损失，以及总体指标——GLUE/MRPC 中的 F1分数和准确率。

表10中，五种基于 BERT 的预训练模型在冻结层并执行相似下游任务时，训练损失和验证损失介于 20% 至 45% 之间。该领域的多标签分类任务有两个变量，具体来说，输入为句子1和句子2，输出则是用于判断相似性或不相似性的二元分类。

验证和测试结果显示，这五种预训练模型均能模拟领域多标签分类任务，其表现从好到差依次为 BERT-base-uncased、Legal-BERT- FP、Zlucia/legalBERT、Zlucia/custome-legalBERT,

模型	训练轮次	训练损失	验证损失	验证 准确率	验证 F 分数	测试 准确率	测试 F 分数
BERT-BASE-UNCASED	1	30.71	26.99	90.92	66.99	90.95	67.42
	2	25.53	29.19	91.69	68.30	92.00	69.66
	3	21.44	29.24	92.00	71.67	92.61	74.01
LEGAL-BERT-FP	1	33.97	29.80	88.65	51.10	88.36	49.63
	2	28.49	29.00	91.73	66.95	91.53	66.42
	3	25.21	27.31	92.07	70.46	92.17	71.10
LEGAL-BERT-SC	1	43.17	42.81	84.52	0.00	84.46	0.00
	2	41.17	43.58	84.53	0.17	84.45	0.00
	3	42.25	42.73	84.53	0.17	84.45	0.00
ZLUCIA/CUSTOM-LEGALBERT	1	40.81	38.93	85.04	42.60	84.48	40.34
	2	35.42	35.68	85.75	46.47	85.42	45.26
	3	34.24	35.26	85.19	49.42	85.08	48.83
ZLUCIA/LEGALBERT	1	35.50	34.90	88.48	44.42	88.10	43.40
	2	31.24	31.77	89.36	58.15	89.40	58.32
	3	27.65	31.49	90.08	60.11	90.32	61.51

表10：相似度下游任务的领域多标签分类验证和测试结果（%）

以及 Legal BERT-SC。持续预训练优于从头训练。BERT-base-uncased 表现最佳，而 Legal BERT-SC 表现最差，在验证和测试中的 F 分数接近 0%。Legal-BERT-FP 的表现优于 Zlucia/legalBERT。

总之，在微调与预训练和持续预训练任务的任务分布不同的下游任务时，BERT-BASE 在学习领域分布和任务分布方面具有优势。然而，领域适配的 BERT 的优势并不明显。对于不同的下游任务结构，不同的领域预训练 BERT 模型表现各异。很难判断哪种领域 BERT 优于其他模型，也难以探究这些情况背后的原因。其中一个可能的原因是，负向任务迁移使微调中的学习变得困难。

比较四种领域适配的 BERT 后发现，即使经过领域适配，若预训练模型未在 BERT-BASE-UNCASED 的预训练知识基础上进行持续预训练，仍难以通过微调学习下游任务。

当下游任务更加复杂、充分利用预训练模型中的参数时，微调过程中的训练损失和验证损失都较低，并且能够很好地学习任务分布。此时，领域持续预训练的优势在 Legal-BERT-FP 中有所体现，其表现优于 BERT，但在 Zlu- cia/legalBERT 中并未出现这种情况。而采用所提出的分类指标后，这些优势不再明显。持续预训练的有效性应在进一步的研究中得到验证。

6 结论与未来工作

分类问题不同于抽取问题。抽取任务侧重于边界检测，而分类问题更像是一个选择问题。在类别列表中插入逗号以分隔类别后，类别边界便很容易确定。因此，只要预测结果包含属于正确类别的词元，模型就正确地选出了该类别。

(Karimi 等人，2021年)指出，在输入序列中插入标点符号是一种更简单的数据增强方法，且不会丢失输入信息。这种方法有助于改变句子中词语的位置，同时不打乱词语顺序，因此展现出更强的泛化能力。然而，未来的研究应关注如何选择合适的符号，例如使用输入文本中未出现过的独特符号。如果使用上下文符号，编码过程将生成不准确的数据。

然而，即使使用不正确的构建数据，微调后的模型仍会得出相似的结论。经检查，错误记录的数量很少。预训练模型是否有能力忽略错误输入或与任务无关的输入，仍是一个问题。

在此前的综述撰写中，有观点认为，持续预训练有助于将 Transformer 模型收缩到特定的领域空间中，从而促进下游领域特定任务的微调。尽管实验表明，Legal-BERT 在微调中的表现优于 BERT，

但由于两种模型的评估结果较为接近，因此仍难以解释其工作机制。持续预训练的优势仍不明确。未来的研究希望阐明其中的原因。

显然，使用特定任务的数据进行微调，可以使 BERT 和 Legal-BERT 同时获得领域知识和任务知识。然而，如果不为下游任务恰当地选择预训练模型，就会发生负迁移。因此，需要采取方法判断某个预训练模型是否适合用于下游任务的微调。下游任务实现方案的选择与提示微调研究密切相关。我们将进一步探讨这一主题，结合“如何选择预训练模型”这一问题，回答“如何为下游任务选择提示”。未来的工作还将探索用于作出判断的方法，以及如何判断任务的简单与复杂程度。如验证和测试结果所示，Legal BERT 的表现并不如预期。因此，后续工作将探究领域适配预训练模型表现未达预期的原因。

预测错误的原因可能是标题未包含足够的概念信息来支持领域标签，因为标题就像法律文档的截断版本；我们曾在一份讨论长输入问题的先前实验报告中探讨过这一点。

先前的实验报告提到，标签具有语义。丰富标签知识是有益的，例如，丰富领域驱动标签的领域知识。EUROVOC表明，领域节点的兄弟节点可以作为丰富领域知识的来源。因此，后续实验将利用EUROVOC层级结构，选择层级信息/节点信息作为补充知识，旨在区分不同领域。

本实验采用层冻结来轻量化模型。未来工作将探索其他压缩方法，例如注意力头蒸馏。

构建提示数据时数据量增大的原因，在于需要生成负样本。抽取任务会在同一个输入中同时生成正向答案和负向答案。因此，抽取任务更易于训练和验证。然而，如果缺少合适的下游任务，或者输入过长，抽取任务就无法达到预期效果。因此，实验需要在训练复杂度与合适的下游任务设计之间取得平衡。

抽取出的答案高度依赖于类别在类别列表中的位置。对于多答案问答任务而言，即使标题和领域-树相同，若随机重新排列领域树中的标签，抽取出的答案也会有所不同。在分类中，标签的顺序并不重要。（Chalkidis 等人，2019）Ko等（2020）也提到了抽取式问答中的位置偏差。未来工作需要考虑如何处理用于分类的多答案问答任务中的位置偏差。

(Jurafsky 和 Martin）指出，问答任务存在基于信息检索和基于知识两种范式。在基于信息检索的方法中，问题类型及其他信息有助于系统进行检索。而基于知识的问答则侧重于将查询映射到结构化数据（例如数据库），以便从数据库中检索相关信息。在该系统中，答案就像用户界面与系统数据库之间的接口。与登录系统中的用户验证界面类似，答案接口可以是 SQL代码，自动将用户查询映射为从数据库中检索信息的请求。因此，未来的工作将着重于将领域添加为一种问题类型，并探索用于问答的语义解析器。

致谢

我们感谢所有作者阅读并批准最终稿。

参考文献

Ilias Chalkidis、Manos Fergadiotis、Prodromos Malakasiotis、Nikolaos Aletras 和 Ion Androutsopoulos。2020年。Legal-BERT：《The Muppets Straight Out of Law School》。*arXiv* 预印本 *arXiv:2010.02559*。

IliasChalkidis、ManosFergadiotis、ProdromosMalakasiotis 和 Ion Androutsopoulos。2019。欧盟立法的大规模多标签文本分类。*arXiv*预印本 *arXiv:1906.02192*。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向变换器预训练。*arXiv*预印本 *arXiv:1810.04805*。

Daniel Jurafsky 和 James H Martin。《语音与语言处理：自然语言处理、计算语言学和语音识别导论》。

Akbar Karimi、Leonardo Rossi 和 Andrea Prati。2021年。Aeda：一种更简单的文本分类数据增强技术。*arXiv*预印本 *arXiv:2108.13230*。

Jaejun Lee、Raphael Tang 和 Jimmy Lin。2019。Elsa 会怎么做？在 Transformer 微调期间冻结层。*arXiv* 预印本 *arXiv:1911.03090*。

Shervin Minaee、NalKalchbrenner、Erik Cambria、Nar- jes Nikzad、Meysam Chenaghlu 和 Jianfeng Gao。2021年。基于深度学习的文本分类：全面综述。*ACM Computing Surveys*（*CSUR*），54(3):1–40。

Elad Segal、Avia Efrat、Mor Shoham、Amir Globerson 和 Jonathan Berant。2019。一种简单有效的多跨度问题回答模型。*arXiv* 预印本 *arXiv:1909.13375*。

Carlos N Silla 和 Alex A Freitas。2011。跨不同应用领域的层次分类综述。数据挖掘与知识发现，22:31–72。

Lucia Zheng、Neel Guha、Brandon R Anderson、Peter Henderson 和 Daniel E Ho。2021年。预训练何时有帮助？评估自监督学习在法律领域的应用，以及包含 53,000+ 项法律保留裁定的 CaseHOLD 数据集。载于 第十八届国际人工智能与法律会议论文集，第 159–168 页。