生成模型在层次结构极限多标签文本分类中的应用

陈林卿*，何大望，肖燕思，刘依林，陆剑平，王为磊
（智慧芽信息科技有限公司，江苏 苏州 215000）
{chenlinqing,hedawang,xiaoyansi,liuyilin,lujianping,wangweilei}@patsnap.com

摘要

层次结构极限多标签文本分类是自然语言处理研究领域中一个重要而又具有挑战性的课题。该任务类别标签数量巨大且自成体系，标签与标签之间还具有不同层级间的依赖关系或同层次间的相关性，这些特性进一步增加了任务难度。该文提出将层次结构极限多标签文本分类任务视为序列转换问题，将输出标签视为序列，从而可以直接从数十万标签中生成与文本相关的类别标签。通过软约束机制和词表复合映射在解码过程中利用标签之间的层次结构与相关信息。实验结果表明，该文提出的方法与基线模型相比取得了有意义的性能提升。进一步分析表明，该方法不仅可以捕获利用不同层级标签之间的上下位关系，还对极限多标签体系自身携带的噪声具有一定容错能力。

关键词： 极限多标签文本分类 ； 层次结构极限多标签 ； 生成模型

Generation Model for Hierarchical Extreme Multi-label Text Classification

CHEN Linqing, HE Dawang, XIAO Yansi, LIU Yilin, LU Jianping, WANG Weilei
(PatSnap Co., LTD. Suzhou, Jiangsu 215000)

Abstract

Hierarchical extreme multi-label text classification task is an important yet challenging task in Natural Language Processing. This task is complex due to the enormous number of labels and corresponding hierarchy relationships in the label system. We propose to view the hierarchical extreme multi-label text classification task as a generation problem and present a novel soft-constrained method for label decoding, which views the output labels as a sequence, rather than as a single label. Rigorous experiments demonstrated that our model is effective at picking out relevant labels directly from thousands of hierarchical labels. Experiments also show that the proposed methods have achieved significant improvements across several datasets. With further analysis, our methods not only capture and utilize the hierarchical structure information between labels at different levels, but also represent the relationships of the labels within the same level.

Keywords: extreme multi-label text classification , hierarchical labels , generation model

*Corresponding author

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。
(c) 2022 中国中文信息学会计算语言学专业委员会

1 引言

多标签文本分类利用自然语言处理高效归纳海量文本信息：给定输入文本，从标签集合中返回与输入文本最相关的标签子集。可以将多标签文本分类问题看作学习评分函数的过程，该函数将(实例，标签)对(x, y)映射到分数$f(x, y)$。函数f在模型训练过程中不断优化，使高度相关的(x, y)对获得高分，而不相关的配对得分较低。具有数千甚至更多类别标签的文本分类任务被称为极限多标签文本分类(Liu et al., 2017)。许多现实世界的应用场景都可以看作这种形式。例如，在开放领域的问答中，x代表一个问题，y代表一个包含答案的文章 (Chang et al., 2020; Lee et al., 2019)。相应的，在层次结构极限多标签文本分类（HXMC, Hierarchical EXtreme Multi-label text Classification ）任务中，x表示文本，y表示具备层级结构的标签体系中的一个或多个类别标签。现实应用场景中数据产生速度快，体量大，具有明显的多样性和复杂性。与之对应的标签数量可能高达数万甚至数十万。极限多标签文本分类在档案文献管理，文本资料分类检索等场景具有广泛的应用前景。

极限多标签文本分类任务类内类间样本关系复杂，导致标签语义存在部分重叠并非完全正交。微软发布的学术图谱数据集MAG（Microsoft Academic Graph）(Shen et al., 2018)就是层级结构极限多标签文本分类数据集中的典型。现有研究方法主要有两大类，其中一类利用标签向量压缩 (Liu et al., 2017; Bhatia et al., 2015)等方法减少标签向量维度实现十万级别标签的分类任务，该类方法丢失部分标签语义信息，忽略标签之间的相关性及其层次结构。另一类考虑到标签结构信息的研究工作则多使用硬性约束，通过多次分类，聚类 (Chang et al., 2020)等方法间接或分步骤实现极限多标签分类。忽略了同一文本可以属于多个交叉领域的事实，也没有充分考虑极限多标签体系由于信息量巨大，需要一定的容错能力，若分类/聚类中间步骤产生错误，会一直传播到后续分类结果，形成由错误传递导致的系统性偏见。

本文受到序列到序列(Seq2Seq)模型在机器翻译 (Bahdanau et al., 2014; Luong et al., 2015; Sun et al., 2017)，摘要总结 (Rush et al., 2015; Lin et al., 2018)，风格转移 (Shen et al., 2017; Xu et al., 2018)等一系列序列转换任务上广泛应用的启发，提出利用基于并行多头注意力机制的生成模型来解决层次结构极限多标签文本分类任务。该序列生成模型由带有注意机制的编码器和解码器组成,显著缓解了之前研究工作中CNN感知域太小，LSTM长文本编码能力弱并且编码解码速度慢的缺点，同时多头注意力机制可以分别关注文本的不同部分，保留了CNN多通道输出的优点。解码器基于具备软约束机制的注意力和柱状搜索算法，在之前预测的标签基础上预测下一个标签，挖掘并利用标签序列内部依赖信息。此外，本文提出的标签词表复合映射机制在保留标签体系结构信息的前提下大幅减少词表维度，避免将极限多标签任务拆分为多个分类模型，并确保模型最终输出标签一定存在于标签体系中。

本文主要贡献如下：

- 提出将层次结构极限多标签文本分类任务视为序列转换问题，利用编码器-解码器结构学习类别标签的层级结构关系。
- 提出软约束解码及词表复合映射，在保留标签语义信息的同时为文本从数十万分类中选出高相关标签。而不是将极限分类任务拆分，组合多个分类模型的输出结果作为输出标签。
- 实验表明本文提出的方法与基线模型相比取得有意义的性能提升。进一步的分析表明该方法在学习标签体系层次结构信息方面的有效性。

2 方法

本文利用神经网络自主习得层次结构极限多标签之间的从属，依赖关系，从而通过生成模型为待分类文本匹配多个具有层级依赖关系的类别标签。为了实现这个目标，本文方法在训练过程中将源端编码器自注意力层输出的词级隐藏状态作为文本编码结果。使用解码器对含有软约束层次结构信息的标签序列进行解码。在预测阶段通过柱状搜索预测标签序列，并通过恢复子词化标签及词表映射得到最终输出标签。该章节将详细介绍本文提出的模型的必要细节，为了方便理解还将对一些概念及问题做出定义和解释。

2.1 层次结构极限多标签文本分类

在层次结构极限多标签文本分类任务中。给定具有L个标签的标签空间$\mathcal{L} = \{l_1, l_2, \cdots, l_L\}$，一个包含$m$个字符的文本序列$\mathcal{X} = \{x_1, x_2, \cdots, x_m\}$，目标是将一个与文本高

输入

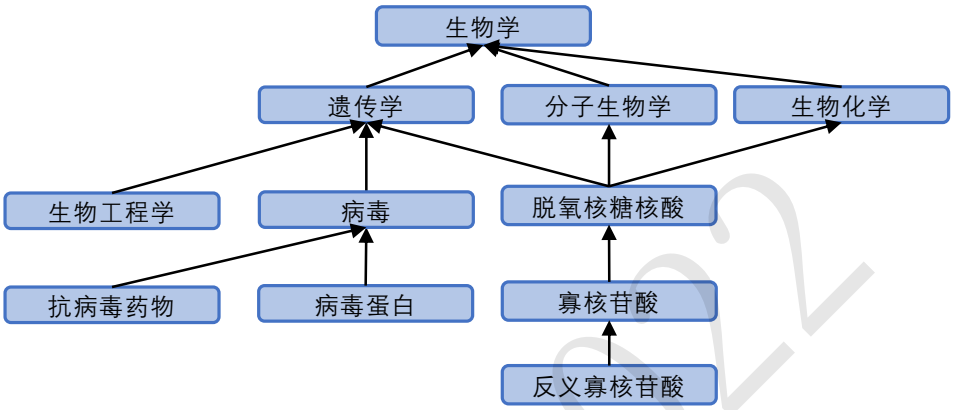
标题：抗冠状病毒的反义寡聚核苷酸及其制药用途

摘要：本发明公开了抗冠状病毒的反义寡聚核苷酸及其制药用途，涉及生物工程领域，解决现有抑制病毒复制的研究大多靶向冠状病毒入侵细胞后被宿主细胞已经翻译生成的RdRp蛋白本身，而不是抑制RdRp的翻译。本发明公开的反义寡聚核苷酸及其联合应用，可特异性地结合冠状病毒5'UTR中IRES序列上的关键茎环结构，寡聚核苷酸序列选自A21，B21，E21和F21等。本发明针对冠状病毒基因组和亚基因组5'UTR进行抗病毒药物设计，是从干扰病毒侵染后多肽段非结构蛋白ORF1ab基因和病毒蛋白翻译的角度考虑，而不是等病毒蛋白大量翻译后再以病毒蛋白作为药物靶点，本发明是以最小的抗冠状病毒成本投入获得最大抗冠状病毒效益产出。

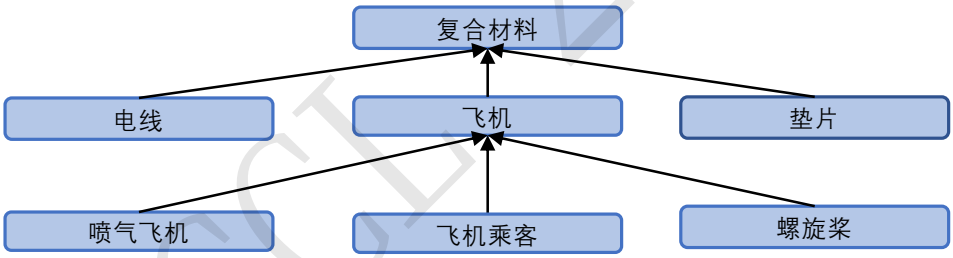
输出

药物；抗病毒药物；病毒蛋白；生物工程学；反义寡核苷酸；基因组；冠状病毒；

(a)：抗病毒生物医药相关专利



(b)：样例专利部分标签的关系示意



(c)：MAG标签体系噪声示例

Figure 1: (a): 样例文本及模型输出的部分标签; (b): 相关标签的层间及层内关系;(c): MAG标签的噪声示例

度相关的$\mathcal{L}$的子集$\mathcal{Y}$分配给文本。与传统的单标签分类只分配给每个文本一个标签不同，每个样例可以有多个标签，且标签之间有一定的从属，依赖关系。标签空间$\mathcal{L}$与文本$\mathcal{X}$之间可以映射n个l。从数学角度来看，HXMC任务可建模为寻找最优标签序列$\mathcal{Y}$的最大化条件概率$p(\mathcal{Y}|\mathcal{X})$问题，其计算公式如下:

$$p(\mathcal{Y}|\mathcal{X}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{P}(\mathbf{y_i}|\mathbf{y_1}, \mathbf{y_2}, \cdots, \mathbf{y_{i-1}}, \mathcal{X}), \tag{1}$$

其中，训练集以待分类文本及标签子集对$(\mathcal{X}, \{y_i\}_{i=1}^{n})$的形式给出。$x_i \in \mathbb{R}^D$，$y_i \in \mathbb{R}^L$。$D$表示文本$\mathcal{X}$中字符$x_i$的特征向量维度,$L$表示标签$y_i$特征向量维度。

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。

MAG[0]是由微软构建的开源异构知识图。如图 1所示，作为层次结构极限多标签体系的典型代表，MAG标签体系具备以下特点：

- 标签数量多，MAG标签体系有约70万个分布在6个不同层次上的类别标签，如图 1(a)所示，一条文本可以有多个高度相关的标签。
- 标签关系复杂，如图 1(b)所示，MAG标签间有层次关系，且一个标签可以从属于多个父节点标签。同层标签之间有一定关联。
- 存在错误信息，由于MAG标签数量高达数十万，其上下位关系甚至标签本身可能存在噪声。图 1(c)展示了其中一个样例，在“复合材料”父节点标签下存在“飞机乘客”子标签，这使得其他研究工作中将极限多标签分类任务按类别拆分成多个分类任务的硬约束方法可能造成错误传递并最终导致系统偏见。

2.2 文本编码与标签生成

该小节介绍G-HXMC模型（Generation mode for Hierarchical EXtreme Multi-label text Classification）的必要细节。本文编码-解码结构基于自然语言处理任务中广泛应用的Transformer(Vaswani et al., 2017)，这里主要介绍涉及本文模型编码，解码的核心部分，其他通用组件如前馈神经网络，残差连接等与经典Transformer相同，因篇幅所限不再详细展开。

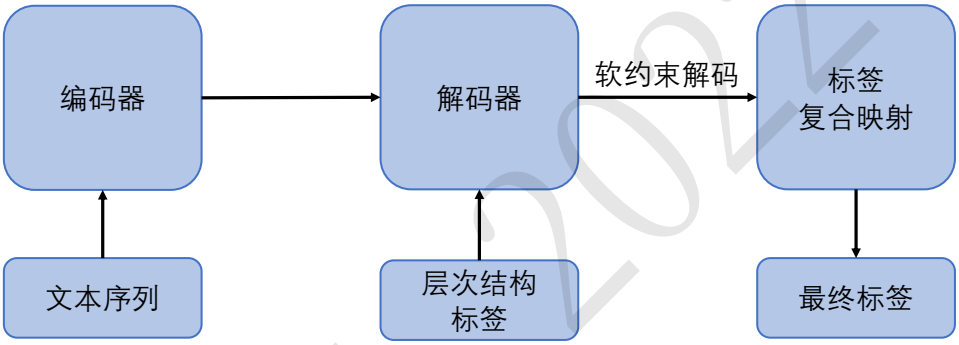


Figure 2: 层次结构极限多标签分类

如图 2 所示，本文提出模型的主要结构由基于多头并行注意机制的编码器和包含软约束及词表复合映射机制的解码器组成。由于模型的生成特性，训练阶段的标签解码与预测阶段的标签生成过程不完全相同。

文本编码 编码器输入由文本标题及摘要拼接成的长序列。与其他序列转换任务类似 (Press et al., 2016)，本文通过词嵌入层将输入序列和输出序列转换为维度D的向量，模型两端的词嵌入层共享同一个线性变换权重矩阵。由于本文长度较长，为了感知输入文本的顺序增加与向量相同维度$\mathbb{R}^{model}$的绝对位置编码(Vaswani et al., 2017)作为位置信息。同时，为了使模型可以感知标题与摘要的区别，本文利用“分段位置编码”区分文本的不同部分。输入序列公式表达如下：

$$I = Concat(PE_1 + emb(T), PE_2 + emb(A)), \tag{2}$$

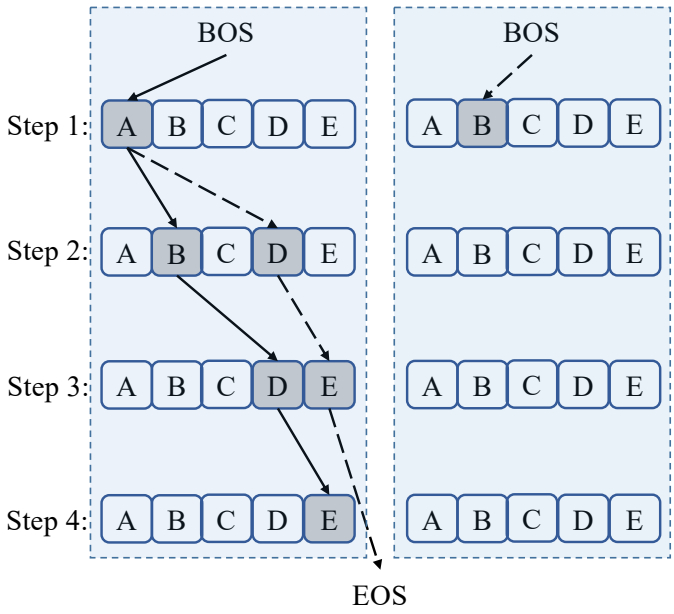
其中，I表示输入序列，T和A分别表示待分类文本的标题与摘要，emb表示词嵌入，PE表示位置编码。

输入文本的不同组成部分并非同等重要，本文使用多头注意力机制捕获不同位置字符之间的关系，对文本进行编码。其公式表达如下：

$$I^{(k)} = MultiHead\left(I^{(k-1)}, I^{(k-1)}, I^{(k-1)}\right), \tag{3}$$

其中，$I \in \mathbb{R}^{model}$表示经过词嵌入的输入序列，K代表编码器层数。

[0]https://www.microsoft.com/en-us/research/project/microsoft-academic-graph



A:生物学;B:遗传学;C:分子生物学;D:生物化学;E:生物工程学

Figure 3: 层次结构标签的生成

标签解码 G-HXMC通过自回归函数:$score(\mathcal{L}|\mathcal{X}) = p_\theta(\mathcal{Y}|\mathcal{X}) = \prod_{i=1}^{N} p_\theta(y_i|y_{i<i}, \mathcal{X})$。其中$\mathcal{Y}$是属于$\mathcal{L}$的具有$N$个子词化字符的标签集合，$\theta$为模型的参数。利用Techer Forcing (Sutskever et al., 2014)最大化输出序列似然度（likelihood），并用dropout(Srivastava et al., 2014)和标签平滑 (Szegedy et al., 2016)进行归一化。简洁地说，本文训练目标即神经机器翻译中的常用目标，通过模型参数θ最大化$logp_\theta(\mathcal{Y}|\mathcal{X})$。

解码器主要组件为两个多头注意力层，一个对标签序列进行编码另外一个利用编码器输出的文本编码结果对标签序列进行解码。解码过程的多头注意力机制公式表达如下:

$$O^{(k)} = MultiHead\left(I^{(k)}, L^{(k)}, L^{(k)}\right), \tag{4}$$

其中，L表示编码后的标签序列，I表示编码后的文本序列，K表示解码器的层数。

标签序列进行编码的方式与公式 3 表示的编码过程类似，但编码内容是目标端标签序列。其他常见组件如前馈神经网络等因篇幅所限亦不再展开。

Cao (2021)等人在具有层次结构的实体召回研究中提出对标签输出过程进行限制。即每一步都进行检查，使得输出结果必然属于前一标签的下位词。强制约束可能存在一些弊端：每个词基于上个输出可能造成错误传递；每一步都进行检查增加时间开销；最重要的是，该约束过程在模型训练过程中并不存在，模型无法学习这种硬约束模式。通过观察数据我们发现，高层级标签天然的具有出现频次高的特性。基于此，本文提出同时在模型的输出及训练过程中对标签进行软约束。根据训练集标签的出现频率对目标端标签进行排序，高层标签排在低层标签前面，同一层标签中高频标签放置在序列前端。

标签生成 极限多标签分类研究工作大多为每条输入文本计算一个维度与标签数一致的输出向量，通过贪心搜索直接为待分类文本选择概率最高的top$N$标签。然而当标签数量很大时，其计算的时间开销和经济成本都十分可观，例如本文使用的MAG数据集有数十万不同标签。同时贪心搜索追求单个位置的最大概率，容错能力较弱，而本文目标是整个序列的概率最大而不是单个标签概率最大。

本文利用介于贪心搜索和广度搜索之间的柱状搜索 (Sutskever et al., 2014)解码策略。如图 3 所示，通过保留一定容错空间缓解贪心搜索可能发生错误传递的缺点。使用柱状搜索的时间开销不取决于词表的大小，只取决于解码过程中保持柱的数量（K）以及解码长度。图中示例K为2，即同时保持2条总体概率最高的候选链路。图 3 中分别以虚线和实现代表两条候选序列实际路径，“Setp2”解码第2个标签时，所有以B标签为第一个标签的候选序列“B-X”概率都小

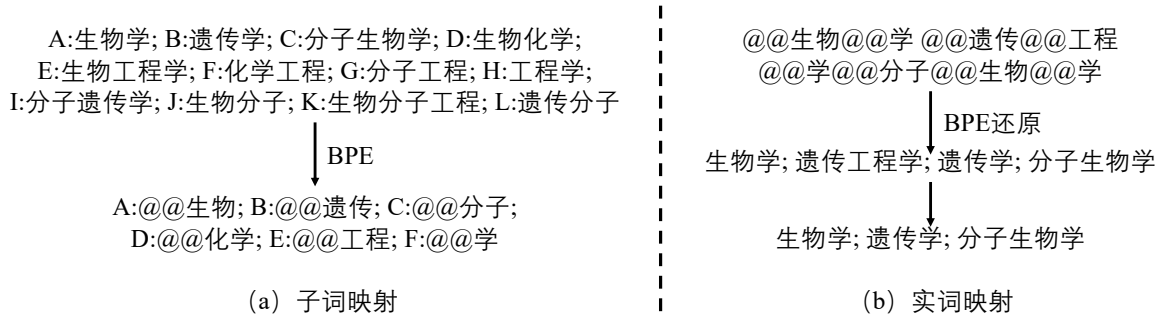


Figure 4: 模型词表与标签的二级映射

数据集	论文			专利		
	训练集	开发集	测试集	训练集	开发集	测试集
样本数	1.97M	5,000	5,000	3.73M	5,000	5,000
标签数	16.90M	42,966	43,206	22.98M	30,797	30,623
平均字符	188.55	190.14	190.34	129.81	131.38	130.91
平均标签	8.54	8.59	8.64	6.15	6.16	6.12

Table 1: 训练集，开发集及测试集的统计信息

于以A标签为第一标签的序列“A-B”和“A-D”，所以右侧柱实际凋零。不同生成路径的长短也有可能不同，虚线路径在生成3个标签后先触发了序列终止符“EOS”。

本文提出的标签词表复合映射主要包括预处理阶段的子词化映射和输出结果阶段的实词映射。受神经机器翻译领域研究工作的启发，本文通过BPE (Sennrich et al., 2016)在保留标签语义信息的前提下缩小解码器端词表。如图 4(a) 所示，子词化后的类别标签数量从约70万减少到不高于3万。如图 4(b) 所示，实词化映射通过筛除不属于MAG体系的标签确保模型输出的最终结果是体系内有实际意义的标签。

3 实验

3.1 数据集

微软学术图(MAG)是一个包含论文和专利等科学出版物，出版物间引用关系，以及作者、机构、期刊、会议和研究领域等信息的开源异构图。基于该知识图谱的数据被用于改善Bing, Cortana, Word和Microsoft Academic的体验。本文实验未使用MAG全量数据，表 1 列举了本文实验使用的专利及论文数据集的部分统计信息，‘M’表示百万，采样方式为随机采样。

3.2 实验设置

本文利用THUMT[1](Tan et al., 2020) 实现基于序列转换的多分类模型，通过拓展词表映射和软约束解码进一步实现层次结构极限多标签文本分类模型。本文在3.5 节列出的实验中，将模型隐藏状态向量的维度设为512，每个编码器解码器的层数都设置为6，多头注意力机制中注意力头的个数都设置为8，柱状搜索的大小设置为5，dropout比例设置为0.1。本文在模型训练过程中将批大小设置为40280个字符并使用$\beta_1 = 0.1$的Adam优化器对模型进行优化 (Kingma and Ba, 2015)。

3.3 评价标准

本文参考之前的研究工作(Wang et al., 2021)，采用微F_1(Micro-F_1)作为主要评价指标。同时报告微准确率（Micro-P）和微召回率（Micro-R）作为辅助参考。Micro-F_1(Manning, 2008)可以理解为Micro-Precision和Micro-Recall的调和平均值。

[1] https://github.com/THUNLP-MT/THUMT

模型	MAG-Paper					
	0-1			0-5		
	m-F	m-P	m-R	m-F	m-P	m-R
#1 G-HXMC	76.38	80.65	72.54	60.64	64.39	57.30
#2 + 软约束解码	80.48	**84.68**	76.68	**63.34**	**67.03**	**59.91**
#3 SGM(Yang et al., 2018)	69.89	69.98	69.80	50.81	63.40	42.39
#4 + Global Emb	70.64	70.69	70.59	50.90	63.78	42.35
#5 + 软约束解码	70.98	71.06	70.90	52.90	65.15	44.53
#6 BERT(Devlin et al., 2018)	78.25	79.38	77.15	——	——	——
#7 + 软约束信息	**81.03**	82.09	**80.00**	——	——	——

Table 2: 本文模型在MAG-Paper任务上的MicroF$_1$性能(%)

模型	MAG-Patent					
	0-1			0-5		
	m-F	m-P	m-R	m-F	m-P	m-R
#1 G-HXMC	64.30	68.34	60.71	57.73	66.75	50.85
#2 + 软约束解码	**71.80**	**76.06**	**67.99**	**64.03**	**68.48**	**60.12**
#3 SGM(Yang et al., 2018)	59.81	64.23	55.96	53.01	39.82	79.25
#4 + Global Emb	60.62	65.03	56.77	53.50	58.01	49.64
#5 + 软约束解码	62.50	67.05	58.53	56.84	61.37	52.93
#6 BERT(Devlin et al., 2018)	63.02	67.14	59.38	——	——	——
#7 + 软约束信息	70.69	75.19	66.70	——	——	——

Table 3: 本文模型在MAG-Patent任务上的MicroF$_1$性能(%)

3.4 基线模型

- BERT(Devlin et al., 2018)，利用基于Transformer的双向编码器进行预训练。预训练后的BERT作为语言模型与可以额外业务层结合广泛应用于下游任务，如问答和语言推断。
- SGM(Yang et al., 2018)，使用基于LSTM的序列转换模型解决极限多标签文本分类。该模型记忆之前时间步的输出并加以利用，用于缓解LSTM的遗忘效应。
- HSG(Wang et al., 2021)，提出一种利用层级标签语义信息引导的模型提升策略,在训练和预测过程中给予模型弱监督语义指导信息,从而规约对应的多标签语义边界。

3.5 实验结果

G-HXMC表示本文基于生成范式的层次结构极限多标签文本分类方法。**软约束解码**表示前文提到的对解码端标签进行排序并结合词表复合映射的方法。**SGM**的实验结果中，**Global Embedding**表示相关论文中使用之前时间步解码结果缓解LSTM遗忘信息的方法。本文作者对该模型进行拓展，使其可以应用本文提出的软约束解码，并报告了相关实验结果。**BERT**不具备解码器结构，无法直接从数十万标签中选出与文本高相关的结果，未报告其0-5层标签体系上的分类实验结果，仅在0-1层约200个标签的范围内进行了实验对比。**软约束信息**指仅对目标端标签序列进行软约束处理，用以验证标签的依赖，从属关系是否会给BERT带来精度增益。**HSG**是一种训练策略而非独立模型，表 5 中对比了G-HXMC及该文方法在**Wiki10-31**数据集上的最佳性能。

表 2 中列出本文提出模型及方法在MAG-Paper数据集上的性能结果。其中左侧为0-1层分类结果，右侧为0-5层分类结果。#2，#5，#7中的实验数据表明，本文提出的方法与基线模型相比取得了有意义的性能提升。其中，本文方法在0-1层和0-5层的分类任务上比SGM提高了约10个点，在0-1层分类任务上与使用大规模预料预训练过的BERT相比使用小的多的数据集和训练开销取得了相近性能。#1与#2的对比表明，本文提出的软约束机制给模型性能带来了显著提升。#3与#5，#6与#7的对比表明本文提出的软约束机制不仅可以给本文提出的模型带来性能增益，也可以帮助基准模型提高性能。

表 3 中列出本文提出方法在MAG-Patent数据集上的实验结果。本文提出的模型在所有层级分类任务中都达到了最佳性能。#2，#5，#7中的实验数据表明，本文提出的方法与基线模型相比取得了有意义的性能提升。其中，本文方法在0-1层和0-5层的分类任务上与SGM和BERT的性能比较趋势与MAG-Paper上的实验结果类似，皆表明本文提出的软约束机制不仅可以给本文提出的模型带来性能增益，也可以帮助基准模型提高性能。

模型	材料学	计算机	化学	工程学	生物学	物理学	医药	平均
G-HXMC	73.10	59.64	66.08	56.21	61.00	57.74	63.00	62.41
+ 软约束解码	**75.07**	**63.53**	**69.13**	**59.20**	**65.47**	**61.97**	**65.33**	**65.66**
SGM(Yang et al., 2018)	62.98	43.10	53.18	47.77	46.74	50.15	49.10	50.43
+ 软约束解码	65.10	45.80	55.40	49.90	49.70	52.00	53.00	52.99

Table 4: MAG-Paper Level 0 标签分类任务MicroF$_1$性能(%).

表 4 列出只根据部分0层标签对MAG-Paper数据集进行分类的实验结果，并将本文模型与基准模型的分类性能进行对比。观察可以发现，本文提出的模型及方法比基准模型具备更强的层次结构信息利用能力，在使用软约束机制之后获得的增益更高。本文方法在使用软约束机制利用标签层级信息后Micro-F$_1$平均增加了3.15，基准模型在使用软约束机制利用标签层级信息后Micro-F$_1$评测标准上平均提高了2.56。

4 分析讨论

为了观察本文提出的基于软约束机制和复合词表映射的生成模型是如何提高层次结构极限多标签文本分类质量的，本文在该章对模型输出标签个数，输出标签质量及一些实验设置进行进一步的实验与分析。实验结果表明，本文模型不仅可以直接从数十万标签中选出与文本相关的标签，还可以自主习得标签体系内部的层次结构信息。

模型	m-F	m-P	m-R
G-HXMC	**56.62**	**61.87**	**52.19**
SGM(2018)	49.89	49.42	50.37
HSG(2021)	47.63	45.26	50.26

Table 5: Wiki10-31上MicroF$_1$性能(%).

模型	参数（百万）	时间（小时）
G-HXMC	66.86	36.5
BERT(2018)	109.64	**30.1**
SGM(2018)	**30.99**	50.3

Table 6: 训练时间和模型参数.

4.1 训练时间和模型参数

表 6 中列举了本文模型和基线模型的参数量及训练时间。其中，SGM的模型参数最少，但由于循环神经网络的特性，其训练时间最长。BERT模型的参数达到1.09亿，由于没有解码器，在不考虑预训练时间的前提下训练时间最短。同时由于BERT没有解码器结构，不能直接从数十万标签中选出与输入文本高度相关的标签，经过改造增加编码器可以实现这一目标，这一方法本质上也是生成模型的范畴，但会使得模型十分庞大，性能也并未取得有意义的增益，不在本文讨论范围内。本文提出的模型及方法在模型规模，训练时间开销等方面的表现较均衡。

4.2 层次结构信息利用效果

模型	标签（个）
G-HXMC	7.07
+ 上下位关系	**7.75**
SGM(Yang et al., 2018)	5.20
+ 上下位关系	5.51

Table 7: 模型输出标签序列长度对比.

本文在表 7 中通过观察不同模型增加软约束层次结构信息前后输出标签个数评估模型利用层次结构信息的能力。该分析基于MAG-Paper数据集。G-HXMC在增加标签上下位信息后标签数量平均增加了0.7个，而SGM则增加了0.3个。由于SGM在增加标签层次结构信息之前的输出标签个数也较少，本文作者推测LSTM不善于编码长序列的缺点限制了SGM对标签层次信息和依赖关系信息的利用。如表 1 所示，数据集中每条文本的平均标签数量不少于8个，且标签具备6个层级，SGM模型的输出标签数量意味着平均每个层级只有不到一个标签，输出标签较少无疑对SGM模型的分类性能带来了较大负面影响。

4.3 层次结构信息利用方式

本文在图 5 中对比了MAG-Paper数据集上标签层次结构信息不同利用方式对分类性能

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。

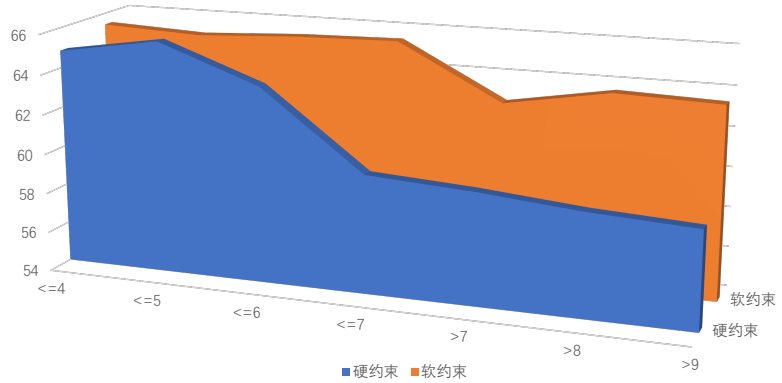


Figure 5: 标签层次结构信息不同利用方式对分类结果的影响MicroF$_1$性能(%).

带来的影响。其中，横坐标表示模型输出标签个数，纵坐标表示对不同长度的结果分别测试MicroF$_1$性能(%)。硬约束表示根据标签层次结构信息对标签从属关系做硬性规定，要求生成的标签必须是前一个标签的下位从属标签。软约束则表述本文所使用的方法。两种利用方式都基于本文提出的G-HXMC模型。通过观察可以看出，硬约束方式输出结果在标签个数较多的时候出现性能大幅下降的现象，而软约束输出结果的性能则较均衡。

4.4 消融实验

标签层级	F1
Level 0-1	71.80
Level 2-5	59.10
Level 0-5	64.03

Table 8: 生成标签层级分布性能对比.

解码长度	F1
5	35.21
10	64.03
15	60.21

Table 9: 不同解码长度对分类性能影响.

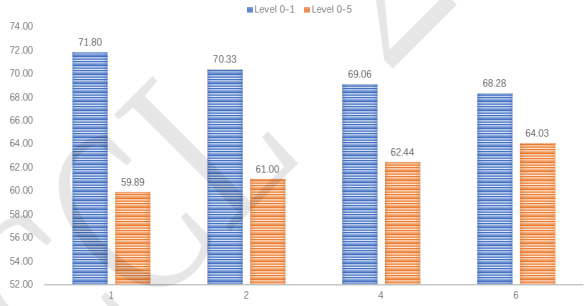


Figure 6: 不同层级分类任务中编码器层数的影响

本文在表 8 中列出了MAG-Patent数据是否利用上位标签信息对分类结果影响的对比。其中，横坐标表示编码器层数，“Level 0-1”指训练数据只有0层和1层标签。“Level 2-5”表示训练数据只有2至5层标签。“Level 0-5”指训练数据包含0-5层的所有标签，但只评估2-5层的分类性能。观察表中结果可知，由于0-1层标签数量较少，标签层次结构清晰无噪声，分类结果总体好于多层次分类结果。另外，对0-5层标签一起学习的分类结果好于只学习2-5层标签的分类结果，表明来自0-1层的优质层次结构信息和依赖关系信息可以给2-5层标签分类带来有意义的增益，验证了本文提出的利用标签体系层次结构信息方法的意义及其有效性。

本文在表 9 中列出了MAG-Paper数据集上，调整解码长度，即输出标签个数对分类结果的影响。表中数据表明使模型输出标签个数与训练数据样本平均标签个数相近可以取得最佳分类效果。这一现象与本文作者的直觉相符。解码长度惩罚系数对模型输出标签个数造成的影响所带来的分类性能变化也体现出类似趋势，由于篇幅所限，本文不再展开介绍。

本文在图 6 中给出编码器层数设置对MAG-Patent数据集不同层级分类任务的影响。其中，“Level 0-1”指只对0层和1层标签进行分类。“Level 0-5”则表示利用标签体系中的所有标签进行分类。观察图中直方图可以发现，0-5层标签分类任务中编码器层数越多性能越好，这一现

象和只对0-1层标签进行分类的实验结果相反。该试验结果可能表明输出标签之间的依赖关系越复杂，生成序列越长，需要的编码器层数也越多。

4.5 样例分析

标题	处理和再生废油产品的方法
摘要	从润滑油或工业油中回收再生矿物油和合成油时..... 然后将搅拌的混合物进行滗析操作，.....
G-HXMC	..... 润滑剂| 石油
+ 硬约束	..... 润滑剂| 油脂	汽车
+ 软约束	..... 润滑剂| 矿物油| 滗析

Table 10: 层次结构信息利用方式样例分析

表 10 中对比了同一专利在不利用标签层级信息及软约束，硬约束两种不同方法利用标签层次结构信息时的输出标签结果。观察样例可以发现，不利用任何标签层级信息时，模型输出标签数量较少。硬性约束的方式利用层次结构信息后虽然输出标签数量增加，但出现了明显不合适的标签“油脂”。软约束利用标签层次结构信息时则给出了专业领域技术标签“滗析”。

标题	具有优良耐磨性和低摩擦系数的钛-石墨烧结复合材料
摘要	.....耐磨性和低摩擦特性的烧结钛-石墨的方法。生产具有可控孔隙率的三相结构的..... .....由于其生物相容性.....该复合材料可用于生物医学工程和其他工程领域.....
硬约束	复合材料| 飞机乘客	石墨| 耐磨
G-HXMC	复合材料| 孔隙率	钛| 烧结| 石墨| 耐磨| 生物相容性

Table 11: 模型输出容错能力样例分析

表 11 中样例对比了不同层次结构信息利用方式对MAG噪声的容错能力。观察样例可以清晰发现，软约束方式不但可以生成更多更贴近专利的标签：“生物相容性”，“孔隙率”等，还绕过了从属于“复合材料”的子标签“飞机乘客”。

5 相关工作

极限多标签文本分类早期工作聚焦基于启发式方法改进传统机器学习方法以适应任务。如Liu等人 (2005)尝试将极限多标签分类任务转化为多个基于支持向量机的二分类问题，这种简洁的策略至今仍被广泛应用；Cai等人 (2004)提出了一种基于支持向量机的层次分类方法来解决极限多标签文本分类任务；SLEEC (Bhatia et al., 2015)通过压缩标签向量维度等辅助手段缓解极限多分类标签文本分类任务中类别标签的“长尾分布”问题。这些研究工作多基于机器学习方法，利用词袋模型对文本语义进行建模。词袋模型忽略词出现的顺序及其之间的语义联系，无法利用上下文，限制了模型理解、分类文本的能力。

神经网络近年来在自然语言处理领域取得了一系列引人注目的成果。一些科研工作者开始在极限多标签任务中应用神经网络模型并取得重要进展。例如,Zhang和Zhou (2006)使用具有成对排序损失函数的全联接神经网络处理功能基因和文本的分类任务。Kurata等人 (2016)提出使用卷积神经网络(CNN)进行多标签分类。Chen等人 (2017)使用卷积神经网络（CNN）和递归神经网络(RNN)来捕捉标签及文本的语义信息。Liu等人 (2017)利用多标签文本分类中经典的深度学习算法如Text Convolutional Neural Networks(TextCNN)、Text Recurrent Neural Networks（TextRNN）、FastText等对文本进行特征抽取从而提升文本语义表示性能。这些研究工作取得了长足进展，但仍未突破卷积神经网络感知域的桎梏以及循环神经网络不善于编码长文本的瑕疵。

在上述工作的基础上，研究者们进一步进行各种优化和改进。Chang等人 (2020)开始尝试利用预训练模型帮助分类任务，但需要利用聚类等机器学习方法对文本进行粗分类；Liu等人 (2017)和Bahatia (2015)则通过压缩标签向量维度降低极限多标签分类任务难度。然而，这些方法大多忽略标签之间的关联，也没有让模型自主习得层次结构标签体系内标签的上下位关系及错综复杂的联系；SGM (Yang et al., 2018)提出利用LSTM对文本进行编码和解码，并通过利用全局信息缓解LSTM不利于长文本编码的弊端。该方法仅在小规模多标签文本分类任务上得到应用，并不适用于具备层次结构的极限多标签文本分类任务；Wang等人 (2021)提出通过标签语义信息加强来改善极限多文本分类任务，将极限多标签分类任务划分为多个分类模型，输出

组合后的分类结果，忽略了错误传递问题，应用于大型深层网络后增益不明显；BERT (Devlin et al., 2018)是最近几年来得到广泛应用的多任务预训练模型。然而该模型不具备解码结构，无法应用于极限多标签分类任务，更无法习得标签之间的联系。Amigo等人 (2022)则拓展了层次结构极限多标签分类任务的评测维度。

6 总结

本文提出将层次结构极限多标签文本分类任务归入生成范式下的序列转换任务。该模型利用编码-解码结构将文本编码后输出一系列标签。为了让模型更好的自主习得类别标签体系的层级结构和依赖关系，本文提出利用软约束解码和复合词表映射帮助模型生成具有层次结构关系的标签序列。在不同类型文本数据集上与多种基线模型的对比实验表明本文提出的一系列方法取得了有意义的性能提升。进一步的消融实验表明，本文提出的软约束机制和复合词表映射方法相比之前的研究工作可以更好的学习利用极限多标签体系的层级结构信息及依赖关系。

如何通过生成模型为文本从数十万甚至上百万具备错综复杂关系的类别标签中选出相关标签，并尽可能多的利用标签间的从属，依赖关系，是一个值得探索并且具备广泛应用前景的问题。我们将在后续工作中继续从不同角度进行有意义的探索并即将分享最新进展，包括但不限于预训练，多模态信息利用，篇章信息利用等。

参考文献

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu and Yiming Yang. 2017. *Deep Learning for Extreme Multi-label Text Classification.* Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 115–124.

Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen and Wei-Ying Ma. 2005. *Support Vector Machines Classification with a Very Large-scale Taxonomy.* ACM SIGIR Explorations Newsletter, 7(1):36-43.

Lijuan Cai and Thomas Hofmann. 2004. *Hierarchical Document Categorization with Support Vector Machines.* Proceedings of the 13th ACM International Conference on Information and Knowledge Management, 78–87.

Kush Bhatia，Himanshu Jain，Purushottam Kar，Manik Varma and Prateek Jain. 2015. *Sparse local embeddings for extreme multi-label classification.* Proceedings of the Neural Information Processing Systems, 29:730-738.

Diederik P. Kingma and Jimmy Ba. 2015. *Adam: A method for stochastic optimization.* Proceedings of ICLR.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu and Houfeng Wang. 2018. *SGM: Sequence Generation Model for Multi-label Classification.* Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 115–124.

Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. *Neural Machine Translation of Rare Words with Subword Units.* Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL, 4171–4186.

Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. *Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising.* WWW.

Yashoteja Prabhu and Manik Varma. 2014. *Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning.* KDD.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. *Pre-training Tasks for Embedding-based Large-scale Retrieval.* International Conference on Learning Representations.

第二十一届中国计算语言学大会论文集，第671页-第683页，南昌，中国，2022年10月14日至16日。

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. *Latent retrieval for weakly supervised open domain question answering.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. *Neural machine translation by jointly learning to align and translate.* CoRR,abs/1409.0473.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. *Effective approaches to attention-based neural machine translation.* CoRR,abs/1508.04025.

Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. 2017. *Label embedding network: Learning label representation for soft training of deep networks.* CoRR,abs/1710.10393.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. *A neural attention model for abstractive sentence summarization.* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 379–389.

Junyang Lin, Xu Sun, Shuming Ma, and Qi Su. 2018. *Global encoding for abstractive summarization.* ACL.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. *Style transfer from non-parallel text by cross-alignment.* CoRR,abs/1705.09655.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. *Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach.* ACL.

Min-Ling Zhang and Zhi-Hua Zhou. 2006. *Multilabel neural networks with applications to functional genomics and text categorization.* IEEE Transactions on Knowledge and Data Engineering, 1338–1351.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. *Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence.* The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 521–526.

Guibin Chen, Deheng Ye, Zhenchang Xing, Jieshan Chen, and Erik Cambria. 2017. *Ensemble application of convolutional and recurrent neural networks for multi-label text categorization.* 2017 International Joint Conference on Neural Networks, 2377–2383.

WANG Yuan, XU Tao, WANG Shilong, ZHOU Yubo and SHI Yancu. 2021. *An Extreme Multi-label Text Classification Strategy via Hierarchical Label Semantic Guidance.* Journal of Chine Information Processing, 35(10):110–118.

Wei-Cheng Chang, Hsiang-Fu Yu and Kai Zhong. 2020. *Taming pre-trained transformers for extreme multi-label text classification.* Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and DataMining, 3163–3171.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2020. *Sequence to sequence learning with neural networks.* Advances in neural information processing systems, 3104–3112.

Ofir Press and Lior Wolf. 2016. *Using the output embedding to improve language models.* preprint arXiv, 1608.05859.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez and Łukasz Kaiser. 2017. *Attention is all you need.* NIPS.

Nicola De Cao, Gautier Izacard, Sebastian Riedel and Fabio Petroni. 2021. *AUTOREGRESSIVE ENTITY RETRIEVAL.* ICLR 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. *Dropout: A simple way to prevent neural networks from overfitting.* Journal of Machine Learning Research, 15(56):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. *Rethinking the inception architecture for computer vision.* Proceedings of the IEEE conference on computer vision and pattern recognition, 2818–2826.

Zhixing Tan, Jiacheng Zhang, Xuancheng Huang, Gang Chen, Shuo Wang, Maosong Sun, Huanbo Luan and Yang Liu. 2020. *THUMT: An Open Source Toolkit for Neural Machine Translation.* AMTA 2020.

Christopher D Manning, Prabhakar Raghavan and Hinrich Schutze. 2008. *Introduction to information retrieval.* volume 1. Cambridge university press Cambridge.

Zhihong Shen, Hao Ma and Kuansan Wang. 2018. *A Web-scale system for scientific knowledge exploration.* ACL 2018.

Enrique Amigo and Augstin D. Delgado. 2022. *Evaluating Extreme Hierarchical Multi-label Classification.* ACL 2022,5809-5819.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK.