NER-guided Comprehensive Hierarchy-aware Prompt Tuning for Hierarchical Text Classification

Fuhan Cai[1], **Duo Liu**[1], **Zhongqiang Zhang**[1], **Ge Liu**[1], **Xiaozhe Yang**[2], **Xiangzhong Fang**[1]
[1]Shanghai Jiao Tong University, [2]East China Normal University
{freyacaifuhan, liuduo, zhangzhongqiang,liu.ge, xzfang}@sjtu.edu.cn
worldetyang@gmail.com

Abstract

Hierarchical text classification (HTC) is a significant but challenging task in natural language processing (NLP) due to its complex taxonomic label hierarchy. Recently, there have been a number of approaches that applied prompt learning to HTC problems, demonstrating impressive efficacy. The majority of prompt-based studies emphasize global hierarchical features by employing graph networks to represent the hierarchical structure as a whole, with limited research on maintaining path consistency within the internal hierarchy of the structure. In this paper, we formulate prompt-based HTC as a named entity recognition (NER) task and introduce conditional random fields (CRF) and Global Pointer to establish hierarchical dependencies. Specifically, we approach single- and multi-path HTC as flat and nested entity recognition tasks and model them using span- and token-based methods. By narrowing the gap between HTC and NER, we maintain the consistency of internal paths within the hierarchical structure through a simple and effective way. Extensive experiments on three public datasets show that our method achieves state-of-the-art (SoTA) performance.

Keywords: Hierarchical Text Classification, Prompt Tuning, Named Entity Recognition

1. Introduction

HTC is a subtask of multi-label classification in NLP where a sample is categorized as a set of labels with a hierarchical structure (Vens et al., 2008). Frequently, real-world classification datasets consist of numerous categories that are organized like this, such as scientific literature categorization (Kowsari et al., 2017) and news corpus (Lewis et al., 2004; Evan Sandhaus, 2008). In these scenarios, the appropriate modeling of the hierarchical structure is a crucial aspect of achieving well-performing classification results, as it allows for an intuitive representation of the complicated relationships among labels.

Inspired by GPT-3 (Brown et al., 2020) and LAMA (Petroni et al., 2019), more and more researches applying prompts (Ding et al., 2022) for model fine-tuning narrows the gap between the pretraining strategies of PLMs and the downstream tasks, and proves that such prompts approach indeed exhibits better performance. Recently, a series of studies have commenced exploring prompt-based learning in the HTC task (Wang et al., 2022b; Ji et al., 2023) and achieved promising results, providing us with more insights to address HTC problems. Despite the success of prompt tuning in HTC, further exploration is needed for its integration with hierarchical structures.

The current SoTA HTC models tend to represent hierarchy features through graph neural networks (GNN) and then inject them as external knowledge into the text features (Wang et al., 2022a; Zhou et al., 2020), leading to notable results. With the rise of pretrained language models (PLMs), an increasing number of tasks are incorporating PLMs (Chen et al., 2021), and HTC is no exception. Moreover, the fusion of hierarchical knowledge has been categorized into three types: injecting features before the model input (Jiang et al., 2022), after the model output (Ji et al., 2023), and simultaneously

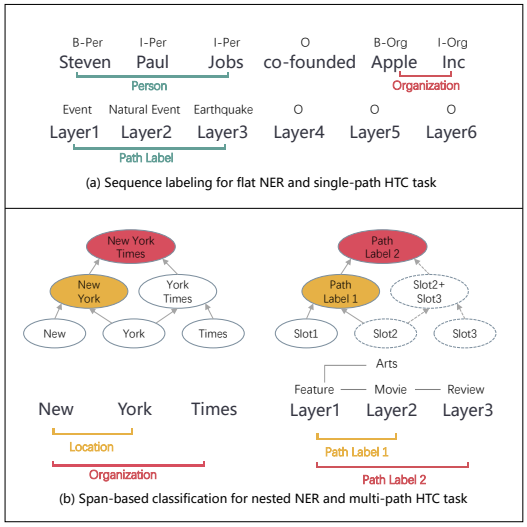

Figure 1: Illustration of method for comparing NER with HTC. (a) For flat NER, each token corresponds to an output label. Similarly, for single-path HTC, each hierarchical slot produces one label. (b) For multi-path HTC, a span-based approach similar to nested NER can be adapted.

LREC-COLING 2024, pages 12117–12126

inserting features (Wang et al., 2022b). Although existing studies have taken into account the positional impact of feature fusion, very few focus on the global and local hierarchical features at the same time, often overlooking the local relationships, specifically, the issue of path consistency within the hierarchical structure. Jiang et al. (2022) notice this issue and propose a sequence-to-sequence model HBGL. However, this method employs pre-extracted global features to represent local ones and uses the sum of embeddings to represent the current level. It fails to capture the specific dependencies among labels within the hierarchy. Besides, additional iterations and sequential constraint prediction are also limitations of their performance.

To address the prompt-based HTC task, which ignores path consistency within the internal hierarchy , we look closely at the label structures and find that the most crucial aspect of the local features lies in the parent-child relationship between adjacent hierarchical levels. Under this observation, we draw parallels with NER tasks that also exhibit constrained relationships among labels, such as the BIO tag scheme (Huang et al., 2015), where the B-tag must follow a start or O-tag, and the I-tag must follow a B-tag, similar to how sub-labels at different levels must follow their corresponding parent labels. Furthermore, as shown in Figure 1, we also identify similarities between single- and multi-path HTC tasks with respect to flat and nested NER problems. In this work, we introduce an NER-guided comprehensive hierarchy-aware model combined with the prompt tuning method to enable prompt-based models to establish complete hierarchical dependencies. Our main contributions are as follows:

- By considering the characteristics of hierarchical dependencies, we are the first to adopt NER methods for modeling two prompt-based HTC tasks, providing a novel perspective for hierarchical-related work.
- We employ Global Pointer and CRF designed for nested and flat entities to model both multi-path and single-path HTC problems, ensuring path consistency in the results through a simple and effective way.
- We evaluate our method on three popular datasets: Web-of-Science (WOS), NYTimes (NYT), and RCV1-V2. Extensive experiments demonstrate that our method achieves significant improvements.

2. Related Work

HTC is a subtask of multi-label classification, which poses challenges owing to its imbalanced, large-scale, and complex hierarchical structures (Mao et al., 2019). Current research is primarily focused on addressing HTC issues by integrating hierarchical features as extra knowledge into the text or model. Specifically, these works are grouped into local and global approaches based on their treatment of label structures (Zhou et al., 2020). The local approaches apply classifiers for each node or layer to acquire specific hierarchical representations. Early works on HTC commonly employ local approaches (Wehrmann et al., 2018; Shimura et al., 2018; Banerjee et al., 2019). However, the number of local classifiers varying with the label structure makes local methods less scalable, leading to the prominence of global approaches, which gradually become mainstream.

Global approaches leverage one classifier to model the label hierarchy. Several works employ popular approaches to integrate hierarchical global information, such as attention mechanism (Zhang et al., 2022), meta-learning (Wu et al., 2019), and reinforcement learning (Mao et al., 2019). Subsequently, Zhou et al. (2020) propose that employing a holistic encoder to represent hierarchical relationships can improve performance. From then on, the focus of HTC research gradually shifts from how to represent hierarchical structures to how to effectively integrate hierarchical features into text and models. HyperIM (Chen et al., 2020) and HiMatch (Chen et al., 2021) adopt the approach of projecting label hierarchy and text semantics into a joint embedding space for further presentation. Zhao et al. (2021) propose a self-adaption semantic awareness network to integrate text and label information. Ji et al. (2023) fuse label hierarchy knowledge into verbalizers in prompt few-shot HTC tasks. Besides, several studies (Wang et al., 2022a,b; Jiang et al., 2022; Zhu et al., 2023) jointly model hierarchical features and text, rather than separately representing them and then integrating, resulting in promising outcomes.

3. Preliminaries

3.1. Traditional HTC

As a subtask of multi-label text classification, HTC primarily differs in the unique composition of its label hierarchy. We predefine the label structure as a special directed acyclic graph (DAG) $\mathcal{H} = (Y, E)$, where Y is the complete label set, and E represents the relationships among labels. Typically, $\mathcal{H}$ is a tree-like structure, wherein each node, except for the leaf nodes, contains one or more children. In HTC, given an input text $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, where n is the sequence length, the objective is to predict a multi-label set $\mathbf{y} \subseteq Y$, corresponding to one or more paths in $\mathcal{H}$ starting from the root node.

3.2. NER-guided Prompt-based HTC

We employ an NER-guided approach on top of the prompt-based HTC. Taking the soft prompt as an example, we make the template like this: `[CLS]` **x** `[SEP] [T1] [Hie1] [T2] [Hie2]` ... `[TL] [HieL] [SEP]`, where `[Hie]` slots indicate the labels at each hierarchical level, virtual template words `[T]` are continuously updated during the training process and `L` represents the maximum number of hierarchical levels. For single-path HTC, we employ the sequence tagging method commonly used in flat NER. Let $\mathbf{x}_{\text{hie}} = [x^1_{\text{hie}}, x^2_{\text{hie}}, \ldots, x^L_{\text{hie}}]$ represent the input sequence of hierarchy slots and $\mathbf{y} = \{y_1, y_2, \ldots, y_L\} \in P$ be the output label path, where P is the path set in $\mathcal{H}$. We aim to build a conditional probability model $Pr(\mathbf{y} \mid \mathbf{x}_{\text{hie}})$. For multi-path HTC, we view it as the nested NER and employ a span-based approach. Let $S = \{s_1, s_2, \ldots s_m\}$ be the possible spans. Each span s is denoted as $s[i:j]$, where i and j are the indices of $\mathbf{H_{hie}}$. Our goal is to identify all $s \in P$.

4. Methodology

This section will introduce the proposed comprehensive hierarchy-aware structure in detail. As shown in Figure 2, our framework is divided into two parts: global and local. In the global method, we draw inspiration from the architecture of HPT (Wang et al., 2022b), which we believe effectively captures the global characteristics of the hierarchical structure but overlooks the local hierarchical dependencies. In the local part, we address the issue of path consistency separately. For the multi-path task, we treat it as a nested NER problem and employ Global Pointer (Su et al., 2022a) for resolution. As for the single-path task, we identify its connection with flat NER and utilize a simple method of CRF(Huang et al., 2015) to capture the relationships between adjacent hierarchical labels.

4.1. Global Hierarchy-aware Structure

We adopt the layer-wise soft prompt template from HPT (Wang et al., 2022b), as described above, denoted as: `[CLS]` **x** `[SEP] [T1] [Hie1] [T2] [Hie2]` ... `[TL] [HieL] [SEP]`. In this context, `[T1]` to `[TL]` are virtual template words, and `[Hie1]` to `[HieL]` are special slots for predicting hierarchical labels, where `L` represents the maximum number of levels in the current dataset.

We utilize BERT (Devlin et al., 2019) as text encoder. Initially, embeddings are obtained for each token:

$$\begin{aligned}\mathbf{Emb} &= [\mathbf{X}; \mathbf{T}] \\ &= [\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{t}_1, \mathbf{e}^1_{hie}, \ldots, \mathbf{t}_L, \mathbf{e}^L_{hie}]\end{aligned} \tag{1}$$

where $\mathbf{X}$ denotes word embeddings from $\mathbf{x}_1$ to $\mathbf{x}_n$ and $\mathbf{T}$ is prompt embeddings, which consisting of virtual template embeddings $\{\mathbf{t_i}\}^L_{i=1}$ and hierarchy prediction embeddings $\mathbf{e}_{hie}$. We feed these embeddings into BERT encoder to get corresponding hidden states:

$$\begin{aligned}\mathbf{H} &= \text{BERT}(\mathbf{Emb}) \\ &= [\mathbf{h}_1, \ldots, \mathbf{h}_n, \mathbf{h}_{t_1}, \mathbf{h}^1_{hie}, \ldots, \mathbf{h}_{t_L}, \mathbf{h}^L_{hie}]\end{aligned} \tag{2}$$

where $\mathbf{h}^i_{hie}$ represents the hidden state of the i-th layer, which is generally filtered by verbalizer for subsequent label prediction. We redefine it as:

$$\mathbf{H}_{\text{hie}} = [\mathbf{h}^1_{\text{hie}}, \mathbf{h}^2_{\text{hie}}, \ldots, \mathbf{h}^L_{\text{hie}}] \tag{3}$$

It is confirmed that the hierarchical injection method used in the previous works effectively captures the global features. Following HPT (Wang et al., 2022b), we employ graph attention network (GAT) (Kipf and Welling, 2017), a widely used model for extracting graph structural features. Given a node v in the l-th layer of GAT, the information it can obtain in the $(l+1)$-th layer is defined as:

$$\mathbf{G}^{(l+1)}_v = \text{ReLU}\Big(\sum_{u \in \mathcal{N}(v) \bigcup \{v\}} \frac{1}{c_v} \mathbf{W}^{(l)} \mathbf{G}^{(l)}_u\Big) \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{d_m \times d_m}$ is a learnable parameter matrix, $\mathcal{N}(v)$ represents the set of neighboring nodes of v, and c_v is a normalization constant. To integrate hierarchical features, $\mathbf{t}_i$ in Equation (1) is updated as $\mathbf{t'}_i$:

$$\mathbf{t}'_i = \mathbf{t}_i + \mathbf{G}^K_{t_i} \tag{5}$$

where K is the number of layers in GAT and the neighbors for $\mathbf{t}_i$ are labels in the current layer.

4.2. Local Hierarchy-aware Structure for Multi-path HTC

As mentioned above, we consider multi-path HTC as the nested NER task and employ the Global Pointer (Su et al., 2022a) model to address it. Given that NER tasks emphasize relationships between adjacent tokens or spans, we apply them to prompt-based HTC in order to capture the local hierarchical dependencies. According to Equation 3, we obtain the hidden state $\mathbf{h}_{hie}$ of each $\mathbf{e}_{hie}$, which corresponds to the label hierarchy at each layer. The length of $\mathbf{H}_{\text{hie}}$ is L, and it is known from the Preliminaries section that there are m possible spans, where $m = L(L+1)/2$. To compute the span representation, we first analyze a specific label path $\alpha \in P$ and pass $\mathbf{h}^i_{hie}$, $\mathbf{h}^j_{hie}$ through two feedforward layers:

$$\begin{aligned}\mathbf{q}_{i,\alpha} &= W_{q,\alpha}\mathbf{h}^i_{hie} + b_{q,\alpha} \\ \mathbf{k}_{j,\alpha} &= W_{k,\alpha}\mathbf{h}^j_{hie} + b_{k,\alpha},\end{aligned} \tag{6}$$

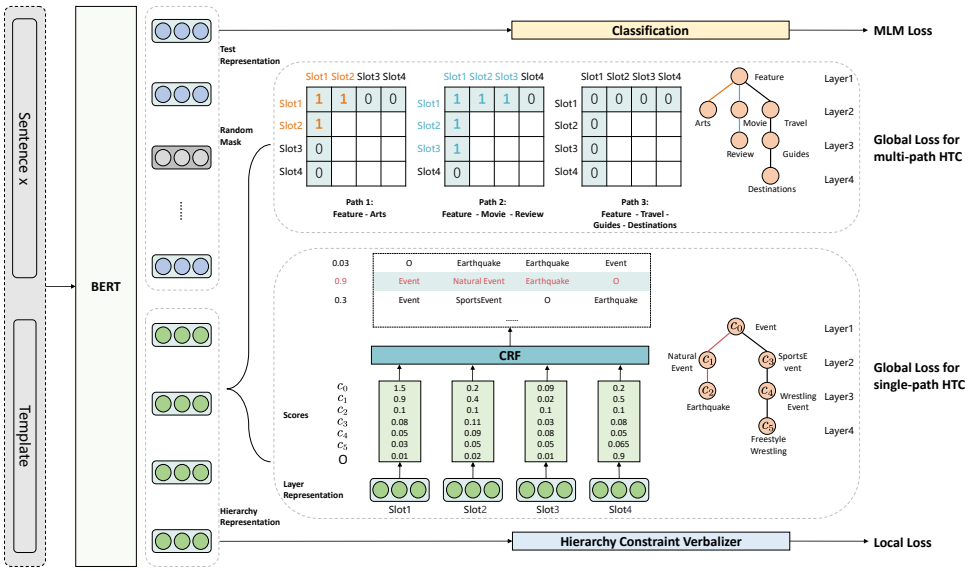

Figure 2: Model structure used in our method. We consider single-path HTC and multi-path HTC as flat and nested NER, respectively. For the multi-path task, we employ span-based Global Pointer method. In the case of single-path HTC, it is treated as a sequence labeling problem. For global-level information, we follow the approach in HPT.

corresponding to the begin and end of the span, respectively. $q_{i,\alpha} \in \mathbb{R}^{d_m}$ and $k_{j,\alpha} \in \mathbb{R}^{d_m}$ are two different vector representations of the hierarchy feature h_{hie} which used to identify the label path α. That is to say, for span $s[i:j]$ of α, $q_{i,\alpha}$ and $k_{j,\alpha}$ represent the start and end position i and j. All vectors in α are denoted as:

$$\begin{aligned} Q_\alpha &= [\mathbf{q}_{1,\alpha}, \mathbf{q}_{2,\alpha}, \cdots, \mathbf{q}_{L,\alpha}] \\ K_\alpha &= [\mathbf{k}_{1,\alpha}, \mathbf{k}_{2,\alpha}, \cdots, \mathbf{k}_{L,\alpha}] \end{aligned} \tag{7}$$

Then, the probability of the span $s[i:j]$ belonging to label path α is computed as follows:

$$s_\alpha(i,j) = \mathbf{q}_{i,\alpha}^\top \mathbf{k}_{j,\alpha} \tag{8}$$

For all path labels in P, the final scores are represented as a tensor $\mathbf{Score} \in \mathbb{R}^{L \times L \times p}$, where p is the total path labels. In HTC tasks, label paths usually start from the root node to leaf nodes. Therefore, we can focus our analysis on $S[1:j]$ to capture relevant information while still conserving space. Following Global Pointer (Su et al., 2022a), we also employ the rotation position encoding (ROPE) to leverage the boundary information, where the transformation matrix satisfies $\mathcal{M}_i^\top \mathcal{M}_j = \mathcal{M}_{j-i}$. After the update, the span scores are represented as:

$$\begin{aligned} s_\alpha(i,j) &= (\mathcal{M}_i \mathbf{q}_{i,\alpha})^\top (\mathcal{M}_j \mathbf{k}_{j,\alpha}) \\ &= \mathbf{q}_{i,\alpha}^\top \mathcal{M}_{j-i} \mathbf{k}_{j,\alpha} \end{aligned} \tag{9}$$

4.3. Local Hierarchy-aware Structure for Single-path HTC

For single-path HTC tasks, each layer outputs just one label, corresponding to the sequence labeling method used in flat NER. Let $\mathbf{x}_{\text{hie}} = [x^1_{\text{hie}}, x^2_{\text{hie}}, \ldots, x^L_{\text{hie}}]$ represent the input sequence of hierarchy slots, and $\mathbf{y} = \{y_1, y_2, \ldots, y_L\} \in P$ represent the output sequence, where P denotes the set of all possible output sequences and L is the sequence length, which is also the maximum hierarchy layers. Let

$$Pr(\mathbf{y} \mid \mathbf{x}_{\text{hie}}) = \frac{\exp(\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}))}{\sum_{\mathbf{y}' \in P} \exp(\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}'))} \tag{10}$$

represent the probability of the output sequence $\mathbf{y}$ given the input sequence $\mathbf{x}_{\text{hie}}$.

To compute the Score, we first define the emission matrix and the transition matrix. Let x^i_{hie} and y_i be the i-th token and its corresponding label in the sequence. The emission matrix $\mathbf{Em} \in \mathbb{R}^{L \times p}$ is such that $Em_{i,j}$ represents the score from x^i_{hie} to y_i, while the transition matrix $\mathbf{Tr} \in \mathbb{R}^{(p+2)\times(p+2)}$ is such that $Tr_{i,i+1}$ represents the probability of transitioning from y_i to y_{i+1}. $\mathbf{Tr}$ contains two additional states, start and end. Then, Score can be calculated as:

$$\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}) = \sum_{i=1}^{L} Em_{i,y_i} + \sum_{i=0}^{L} Tr_{y_i, y_{i+1}} \tag{11}$$

where y_0 and y_{L+1} are the two additional states mentioned above.

Furthermore, following previous work, we consider the hidden states outputted by the feedforward layers or encoders as emission probabilities. Therefore, the formula Score can be transformed

into:

$$\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}) = \sum_{i=1}^{L} \boldsymbol{l}_{\text{hie}}^{i}\ [y_i] + \mathbf{Tr}_{y_i, y_{i+1}} \tag{12}$$

where $\boldsymbol{l}_{\text{hie}}^{i} \in \mathbb{R}^p$ is the i-th logit vector of $\mathbf{H}_{\text{hie}}$ (in Equation 3) after passing through a feedforward layer.

It's worth mentioning that flat entities can also be represented in the form of spans, which means that single-path problems can also be addressed using Global Pointer, as it is a general method.

4.4. Objective Function

Our final training loss function consists of three parts, namely masked language model (MLM) loss, global loss, and local loss:

$$\mathcal{L} = \mathcal{L}_{MLM} + \lambda_1 \mathcal{L}_{Global} + \lambda_2 \mathcal{L}_{Local} \tag{13}$$

where λ_1 and λ_2 are two hyperparameters for balancing the global loss and local loss. To maintain consistency with the pretraining task, we randomly mask 15% tokens and use the $\mathcal{L}_{MLM}$ for training:

$$\begin{aligned} \mathcal{L}_{MLM} &= -\log \frac{e^{s_t}}{\sum_{i=1}^{V} e^{s_i}} \\ &= \log(1 + \sum_{i=1, i \neq t}^{V} e^{s_i - s_t}) \end{aligned} \tag{14}$$

which is also the cross entropy (CE) loss, where s_t is the target score and V is the vocabulary size. Following HPT (Wang et al., 2022b), we also use the Zero-bounded Log-sum-exp & Pairwise Rank-based (ZLPR) loss (Su et al., 2022b) to alleviate the issue of class imbalance in multi-label classification. The original log-sum-exp pairwise (LSEP) loss (Li et al., 2017)

$$\mathcal{L}_{LSEP} = \log(1 + \sum_{i \in \mathcal{U}_{neg}} \sum_{j \in \mathcal{U}_{pos}} e^{s_i - s_j}) \tag{15}$$

cannot handle situations with the variable number of target categories, where $\mathcal{U}_{pos}$ and $\mathcal{U}_{neg}$ are the positive and negative label set. To address this issue, Su et al. (2022b) introduce a threshold s_0 in the loss function, aiming to have the scores of positive classes greater than s_0, and those of negative classes less than s_0. Then, the ZLPR loss function is defined when the threshold s_0 is set to 0 as follows:

$$\begin{aligned} \mathcal{L}_{ZLPR} &= \log(1 + \sum_{i \in \mathcal{U}_{neg}} \sum_{j \in \mathcal{U}_{pos}} e^{s_i - s_j} \\ &+ \sum_{i \in \mathcal{U}_{neg}} e^{s_i - 0} + \sum_{j \in \mathcal{U}_{pos}} e^{0 - s_j}) \\ &= \log(1 + \sum_{i \in \mathcal{U}_{neg}} e^{s_i}) + \log(1 + \sum_{j \in \mathcal{U}_{pos}} e^{-s_j}) \end{aligned} \tag{16}$$

Based on the above equations, the global loss function can be represented as:

$$\begin{aligned} \mathcal{L}_{\text{Global}} = \sum_{m=1}^{L} (&\log(1 + \sum_{i \in \mathcal{U}_m^{neg}} e^{s_i}) \\ &+ \log(1 + \sum_{i \in \mathcal{U}_m^{pos}} e^{-s_i})) \end{aligned} \tag{17}$$

where $s_i = \mathbf{v}_i^T \mathbf{h}_{hie}^m + b_{im}$ and b_{im} is bias. Besides, $\mathcal{U}_m^{pos}$ and $\mathcal{U}_m^{neg}$ are positive and negative label sets for the m-th layer respectively. Similarly, the local loss function for the multi-path task can be represented as:

$$\begin{aligned} \mathcal{L}_{\text{Local}} = \sum_{\alpha \in P} (&\log(1 + \sum_{(i,j) \in \mathcal{Q}_\alpha} e^{s_\alpha(i,j)}) \\ &+ \log(1 + \sum_{(i,j) \in \Omega_\alpha} e^{-s_\alpha(i,j)})) \end{aligned} \tag{18}$$

where i, j are the begin and end indices of a span and $s_\alpha(i, j)$ can be calculated from Equation 8. In addition, Ω_α is a set of spans belonging to path label α and $\mathcal{Q}_\alpha$ is a negative sample set, including spans that are not paths or whose type is not α. When dealing with a single-path HTC task, according to Equation 10, we modify the local loss function to:

$$\begin{aligned} \mathcal{L}_{\text{Local}} &= -\log(Pr(\mathbf{y} \mid \mathbf{x}_{\text{hie}})) \\ &= \log(\sum_{\mathbf{y}' \in P} \exp^{\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}')}) - \text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}) \end{aligned} \tag{19}$$

5. Experiments

5.1. Datasets and Evaluation Metrics

To evaluate the proposed method, we conduct experiments on three widely used datasets for HTC: Web-of-Science (WOS) (Kowsari et al., 2017), RCV1-V2 (Lewis et al., 2004) and NYTimes (NYT) (Evan Sandhaus, 2008). Among them, RCV1-V2 and NYT are for multi-path HTC while WOS include single-path hierarchical labels. The statistical details are shown in Table 1. We follow the data preprocessing methods of previous studies (Zhou et al., 2020; Wang et al., 2022b) and leverage the same evaluation metrics: Macro-F1 and Micro-F1.

Dataset	$\lvert Y \rvert$	Depth	Avg($\lvert y_i \rvert$)	Train	Dev	Test
WOS	141	2	2.0	30,070	7,518	9,397
NYT	166	8	7.6	23,345	5,834	7,292
RCV1-V2	103	4	3.24	20,833	2,316	781,265

Table 1: Data Statistics. $\lvert Y \rvert$ is the number of classes. Avg($\lvert y_i \rvert$) is the average number of classes per sample. Depth is the maximum level of label hierarchy.

Model	WOS (Depth 2)		RCV1-V2 (Depth 4)		NYT (Depth 8)		Average	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Hierarchy-Aware Models								
TextRCNN (Zhou et al., 2020)	83.55	76.99	81.57	59.25	70.83	56.18	78.65	64.14
HiAGM (Zhou et al., 2020)	85.82	80.28	83.96	63.35	74.97	60.83	81.58	68.15
HTCInfoMax (Deng et al., 2021)	85.58	80.05	83.51	62.71	-	-	-	-
HiMatch (Chen et al., 2021)	86.20	80.53	84.73	64.11	-	-	-	-
Pretrained Language Models								
BERT (Wang et al., 2022a)	85.63	79.07	85.65	67.02	78.24	65.62	83.17	70.57
BERT+HiAGM (Wang et al., 2022a)	86.04	80.19	85.58	67.93	78.64	66.76	83.42	71.63
BERT+HTCInfoMax (Wang et al., 2022a)	86.30	79.97	85.53	67.09	78.75	67.31	83.53	71.46
BERT+HiMatch (Chen et al., 2021)	86.70	81.06	86.33	68.66	-	-	-	-
HGCLR (Wang et al., 2022a)	87.11	81.20	86.49	68.31	78.86	67.96	84.15	72.49
HiTIN (Zhu et al., 2023)	87.19	81.57	86.71	**69.95**	79.65	69.31	84.52	73.61
HiTIN†	86.75	81.18	86.72	68.94	79.40	68.42	84.29	72.85
HPT (Wang et al., 2022b)	87.16	81.93	87.26	69.53	80.42	70.42	84.95	73.96
HPT†	86.93	81.50	87.39	69.11	80.59	70.45	84.97	73.69
NERHTC (Ours)	**87.42**$^{\uparrow 0.49}$	**81.93**$^{\uparrow 0.43}$	**87.50**$^{\uparrow 0.11}$	69.76$^{\uparrow 0.65}$	**80.97**$^{\uparrow 0.38}$	**70.99**$^{\uparrow 0.54}$	**85.30**$^{\uparrow 0.33}$	**74.23**$^{\uparrow 0.54}$

Table 2: The experimental results (%) of our proposed method comparing to previous models on three datasets. Best results are in boldface. Our re-implementation scores are marked by "†". "↑" indicates the improvement of our model compared with the second-best result within our reproduction.

5.2. Implement Details

Our model is implemented with Pytorch framework by an end-to-end training style. Following previous work (Wang et al., 2022b), we utilize `bert-base-uncased` pretrained model as the base architecture. The batch size is set to 16. The optimizer is Adam with a learning rate $3e^{-5}$ for BERT and $5e^{-4}$ for single-path model CRF. The loss balancing parameter λ_1 is set to 1. λ_2 is set to 1 on RCV1-V2 and 10 on WOS and NYT. The length of template words for soft prompt is adapted according to the number of layers. We train our model with training set and evaluate on development set. We set the maximum training epochs to 50 and initiate evaluation after training 20 epochs. We employ the following early stopping strategy: training stops when both Macro-F1 and Micro-F1 metrics cease to improve for more than 15 steps. For baseline models, we follow the parameters and settings in the original papers.

5.3. Baselines

We compare our model with multiple HTC baselines as follows:

- **TextRCNN** (Lai et al., 2015): A traditional text classification model, commonly employed as a text encoder before the emergence of large-scale PLMs.
- **BERT** (Devlin et al., 2019): An effective and widely used PLM, often employed as a text feature extractor, capable of capturing semantic information and applying it to various downstream tasks.
- **HiAGM** (Zhou et al., 2020): This hierarchical-aware global model extracts label-wise text features based on prior hierarchy information by hierarchical encoders.
- **HTCInfoMax** (Deng et al., 2021): HTCInfoMax addresses the issues in HiAGM by introducing information maximization, including maximizing text-label mutual information and label prior matching.
- **HiMatch** (Chen et al., 2021): Himatch introduces a hierarchical-aware label semantic matching network, redefining the text-label semantic relationship as a semantic matching problem.
- **HGCLR** (Wang et al., 2022a): HGCLR utilizes contrastive learning to integrate hierarchical features into the text encoder, and introduces a new graph encoder.
- **HiTIN** (Zhu et al., 2023): HiTIN is a simple and efficient architecture in fusing the label structural information into text representations.
- **HPT** (Wang et al., 2022b): HPT adopts prompt tuning to address the HTC problem by incorporating the hierarchical label knowledge into virtual templates and label words.

5.4. Main Results

Table 2 presents the experimental results for each of the models mentioned above. For a fair comparison, we implement some crucial experiments on our own device. As shown, our method outperforms the baseline models on all three datasets. These results demonstrate the superiority of our model since it makes HPT pay attention to local hierarchical dependencies to some extent using an NER-guided method.

On WOS, our model observes 1.79% and 2.86% improvements of Micro-F1 and Macro-F1 respectively against BERT and is better than HPT by 0.46% on average. As mentioned in 4.3, the flat

Ablation Models	Micro-F1	Macro-F1
NERHTC	**80.97**	**70.99**
r.m. ROPE	80.80	70.40
r.m. MLM loss	80.67	70.40
r.p. BCE loss	80.70	70.10
r.m. GAT	80.82	70.25

Table 3: Ablation study results on NYT. *r.m.* stands for `remove`. *r.p.* stands for `replaced with`.

NER problem can be tackled using a span- or token-based approach. In Table 2, to demonstrate a general approach, the results for the single-path dataset WOS are achieved by Global Pointer. In practice, when applying token-based method, the Micro-F1 score is 87.16%, and the Macro-F1 score is 81.63%, which is 0.36% higher than HPT in total.

Among the three datasets, the results for WOS and NYT are favorable. Notably, upon comparison, NYT stands out as the most complex, featuring the highest number of paths, hierarchical levels, and the largest average number of labels. Despite these challenges, we achieve the best results on both evaluation metrics by 0.46% on average. This further underscores the effectiveness of our approach, particularly in addressing intricate multi-path tasks.

6. Analysis

6.1. Ablation Study

We conduct ablation experiments on NYT, as it is the most complex, with the highest number of hierarchy levels and paths. The results are shown in Table 3. After removing positional encodings, we observe a 0.59% decrease in Macro-F1. As the current mainstream choice, RoPE positional encoding (Su et al., 2021) has found increasing application in various large-scale models. We also adopt it to enhance the model's sensitivity to span boundary information. MLM is one of the fundamental tasks in language model pretraining. As a multi-task model, we comprise global, local, and MLM loss. Upon removing MLM loss, we observe a decline in model performance of 0.30% in Micro-F1 and 0.59% in Macro-F1, underscoring the necessity of maintaining relative consistency with the pretraining tasks. In our model, we employ ZLPR multi-label classification loss function. When replaced with the traditional BCE loss, there is a significant decrease of 0.89% in Macro-F1 score, an indicator to evaluate class imbalance issues. From this, ZLPR loss demonstrates better performance in addressing imbalance concerns. In extracting of global features, we employed GAT, which is widely used to represent graph structures. By removing GAT, the model cannot access the information of the label hierarchy and drops 0.74% on Macro-F1.

Method	WOS			
	PMicro-F1	PMacro-F1	CMicro-F1	CMacro-F1
BERT	79.96	78.40	85.43	79.37
HiTIN	81.06	79.23	86.45	80.76
HPT	80.69	79.03	86.57	80.85
NERHTC	**81.41**	**79.52**	**87.14**	**81.36**

Table 4: Consistency experiments of path-based (P-metric) and path-constrained (C-metric) evaluation metrics on WOS.

Method	RCV1-V2		NYT	
	CMicro-F1	CMacro-F1	CMicro-F1	CMacro-F1
BERT	85.68	66.96	78.05	64.62
HiTIN	86.48	68.07	78.45	66.79
HPT	86.95	68.15	79.51	68.38
NERHTC	**86.99**	**68.46**	**80.11**	**69.42**

Table 5: Consistency experiments of path-constrained (C-metric) evaluation metrics on RCV1 and NYT.

6.2. Effect of Local Structure

Our method is proposed to address the limitation of neglecting local hierarchical dependencies in most prompt-based models. To demonstrate the effectiveness of our model, we conduct analyses from three perspectives: path consistency, label granularity, and discriminability.

6.2.1. Path Consistency

We hold the view that path consistency reflects the attention on local features, which often tend to focus more on relationships between adjacent nodes, thereby ensuring the connectivity of paths. Following Yu et al. (2022) and Ji et al. (2023), we utilize path-constrained metric (CMicro-F1 and CMacro-F1) and path-based metric (PMicro-F1 and PMacro-F1) to measure our model's performance at the path level. The C-metrics are defined such that a node label is considered predicted correctly only when all its ancestor nodes are correct. P-metrics measure the correctness of all labels along the entire pathway for mandatory-leaf (Bi and Kwok, 2014) dataset, such as WOS. We re-evaluate the experimental results on three datasets, as depicted in Table 4 and 5.

For the mandatory-leaf dataset WOS, our approach exhibits an overall advantage. It surpasses HiTIN by a total of 1.93% across the four metrics and exceeds HPT by 2.29%. WOS has only two hierarchical levels, which is relatively simpler. On the 8-layer NYT dataset, our approach demonstrates a significant edge, leading HiTIN by 1.66% on CMicro-F1 and 2.63% on CMacro-F1. When compared to HPT, our approach excels on multi-path datasets, particularly outperforming in the CMacro-F1 metric for NYT by 1.04%. From the perspective of path consistency, our method generally outperforms other models, particularly in datasets with

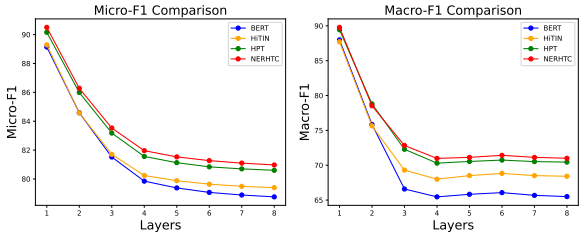

Figure 3: Peference study on label granularity of NYT based on layer increment.

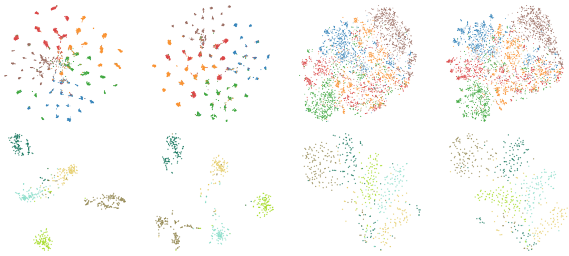

Figure 4: T-SNE visualization of the label representations on WOS. Images in the first row display feature clusters for the first-level labels, while the second row is for the sub-labels of ECE. Dots of the same color belong to the same category.

Models	Acc.
Llama 1-7B	**22.08**
w/ demo	19.05
Llama 2-7B	16.28
w/ demo	16.78

Table 6: The accuracy of using large models for inference on the first-level labels of WOS. *demo* represents `demonstration`. *w/* stands for `with`. The results are the average of three experiments

complex hierarchical structures. This is because it simultaneously considers both global and local hierarchical features, offering a more comprehensive approach to addressing the HTC problem.

6.2.2. Label Granularity

Conducting a granularity analysis on complex hierarchical labels is necessary. In HiMatch (Chen et al., 2021), level-based F1 score analysis is employed. However, calculating scores for labels in each layer separately fails to showcase the advantages of our model. This is because our local architecture places a stronger emphasis on the connections between hierarchy levels. Therefore, we introduce a layer-incremental method: computing the cumulative F1 scores for labels at the current hierarchy and its preceding layers. We conduct the analysis using the NYT dataset with eight levels. As shown in Figure 3, the horizontal axis is the current level, while the vertical axis represents the cumulative F1 scores. The Micro-F1 metric exhibit a smooth decreasing trend, with our model performing optimally at each level. Regarding the Macro-F1, our model maintains a relatively superior performance, only slightly lower than HPT at the second layer. Notably, all four models experience a significant drop at the second level, which we attribute primarily to the increased number of labels and the subsequent surge in sample quantity, leading to a certain performance decline.

6.2.3. Discriminability

As shown in Figure 4, we utilize T-SNE visualization analysis to demonstrate our model's discriminative ability with respect to features. From Figures 4(a) and 4(b), it can be observed that BERT and TiHIN treat WOS as a typical multi-class classification task, where each leaf represents a cluster center, with no connection between labels. Both HPT and our method utilize parent nodes as cluster centers. Features with the same parent should exhibit a relatively distant relationship among sub-nodes while maintaining an overall proximity, as demonstrated in Figure 4(d). Our approach outperforms the HPT method, as reflected in the greater distances between different classes in the graph and the tighter aggregation of features within the same category.

6.3. Large Language Model

Large language models (LLMs) have been garnering increasing attention due to their excellent reasoning abilities. To keep pace with technological advancements and lay the foundation for future work, we conduct tests using the LLaMA-7B models (Touvron et al., 2023a,b). Due to hardware limitations and the cost, we only perform inference on the first-level labels of WOS without fine-tuning. As shown in Table 6, the unguided Llama 1-7B model achieves an average accuracy of 22.08% without any demonstrations. In the current stage, LLMs are still unable to effectively address text classification problems with long texts, multiple labels, and numerous categories. Research based on small models remains valuable. Furthermore, there are still many unexplored aspects of LLMs, awaiting further investigation.

7. Conclusion

We propose an NER-guided prompt tuning model for HTC tasks to address the limitation of neglecting local hierarchical dependencies in prompt-based models. Considering the emphasis on adjacent tokens in NER, we model single- and multi-path HTC tasks using the CRF and Global Pointer methods,

which were initially designed for flat and nested entities. Through our research, a connection is established between HTC and NER, and label consistency in prompt-based HTC tasks is effectively reinforced in a sample end-to-end manner. We empirically verify the effectiveness of our method and achieve the SoTA performance.

Limitations

Based on the results of the above indicators, our method appears to be more effective for data with complex hierarchical structures. However, our method is prompt-based, and more complex hierarchical structures often require a larger number of prompt template words, which may compromise the length of the text. Furthermore, these limitations also exist in LLMs. During zero-shot inference, we provide the model with demonstrations for each level of labels. Detailed and specific demonstrations aid the model in understanding the task, but it consumes the length of prediction samples. Additionally, when dealing with numerous levels of labels, it is often impossible to provide detailed demonstrations, resulting in incomplete model's understanding of the task. We will conduct further research on the above-mentioned issues.

Ethics Statement

All datasets in our paper are public. Our experimental results are obtained using the settings and parameters described in this paper. We ensure the authenticity and reproducibility of our article.

Bibliographical References

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *ACL*, pages 6295–6300.

Wei Bi and James T. Kwok. 2014. Mandatory leaf node prediction in hierarchical multilabel classification. *IEEE Trans. Neural Networks Learn. Syst.*, 25(12):2275–2287.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Boli Chen, Xin Huang, Lin Xiao, Zixin Cai, and Liping Jing. 2020. Hyperbolic interaction model for hierarchical multi-label classification. In *AAAI*, volume 34, pages 7496–7503.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *ACL*, pages 4370–4379.

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip S. Yu. 2021. Htcinfomax: A global model for hierarchical text classification via information maximization. In *NAACL-HLT*, pages 3259–3265.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. Openprompt: An open-source framework for prompt-learning. In *ACL*, pages 105–113.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *ACL*, pages 2918–2933.

Ting Jiang, Deqing Wang, Leilei Sun, Zhongzhi Chen, Fuzhen Zhuang, and Qinghong Yang. 2022. Exploiting global and local hierarchies for hierarchical text classification. In *EMNLP*, pages 4030–4039.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*, pages 2267–2273.

Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving pairwise ranking for multi-label image classification. In *CVPR*, pages 1837–1845.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. In *EMNLP-IJCNLP*, pages 445–455.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP*, pages 2463–2473.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: learning hierarchical category structure for multi-label short text categorization. In *EMNLP*, pages 811–816.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Jianlin Su, Ahmed Murtadha, Shengfeng Pan, Jing Hou, Jun Sun, Wanwei Huang, Bo Wen, and Yunfeng Liu. 2022a. Global pointer: Novel efficient span-based approach for named entity recognition. *arXiv preprint arXiv:2208.03054*.

Jianlin Su, Mingren Zhu, Ahmed Murtadha, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2022b. Zlpr: A novel loss for multi-label classification. *arXiv preprint arXiv:2208.02955*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. 2008. Decision trees for hierarchical multi-label classification. *Machine Learning*, page 185–214.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022a. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *ACL*, pages 7109–7119.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: hierarchy-aware prompt tuning for hierarchical text classification. In *EMNLP*, pages 3740–3751.

Jonatas Wehrmann, Ricardo Cerri, and Rodrigo C. Barros. 2018. Hierarchical multi-label classification networks. In *ICML*, pages 5225–5234.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. In *EMNLP-IJCNLP*, pages 4353–4363.

Chao Yu, Yi Shen, and Yue Mao. 2022. Constrained sequence-to-tree generation for hierarchical text classification. In *SIGIR*, pages 1865–1869.

Xinyi Zhang, Jiahao Xu, Charlie Soh, and Lihui Chen. 2022. LA-HCN: label-based attention for hierarchical multi-label text classification neural network. *Expert Systems with Applications*, 187:115922.

Rui Zhao, Xiao Wei, Cong Ding, and Yongqi Chen. 2021. Hierarchical multi-label text classification: Self-adaption semantic awareness network integrating text topic and label level information. In *KSEM*, pages 406–418.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *ACL*, pages 1106–1117.

He Zhu, Chong Zhang, Junjie Huang, Junran Wu, and Ke Xu. 2023. Hitin: Hierarchy-aware tree isomorphism network for hierarchical text classification. In *ACL*, pages 7809–7821.

Language Resource References

Evan Sandhaus. 2008. *The New York Times Annotated Corpus*. Philadelphia: Linguistic Data Consortium, ISLRN 429-488-225-160-9.

Kamran Kowsari and Donald E. Brown and Mojtaba Heidarysafa and Kiana Jafari Meimandi and Matthew S. Gerber and Laura E. Barnes. 2017. *HDLTex: Hierarchical Deep Learning for Text Classification*.

David D. Lewis and Yiming Yang and Tony G. Rose and Fan Li. 2004. *RCV1: A New Benchmark Collection for Text Categorization Research*.