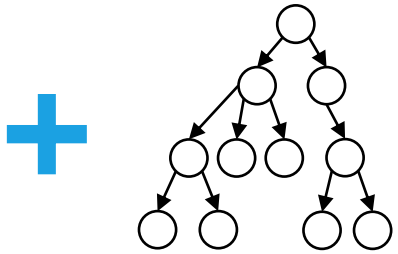
NER-guided Comprehensive Hierarchy-aware Prompt Tuning for Hierarchical Text Classification

Fuhan Cai, Duo Liu, Zhongqiang Zhang, Ge Liu, Xiaozhe Yang, Xiangzhong Fang
Shanghai Jiao Tong University, East China Normal University

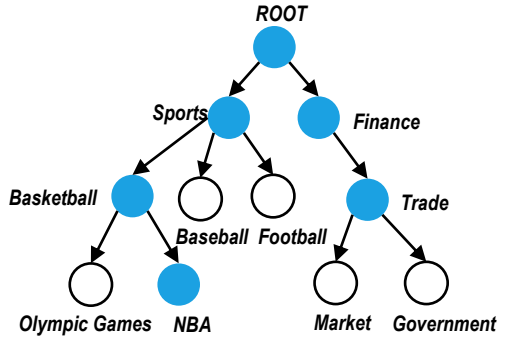
Hierarchical Text Classification (HTC)

Input&Output:

Text: James Harden traded to Clippers.



Label Hierarchy



Labels: Sports, Basketball, NBA, Finance, Trade

Dataset:

- Web of Science (WOS): The input text is the abstracts of articles in the Web of Science citation database , fixed 2-layer, single-path;
- New York Times (NYT): The input text is a news article with up to 8 layers, a multi-path dataset;
- RCV1-V2: The input is Reuters news with up to 4 layers, a multi-path dataset.

Dataset	$\|Y\|$	Depth	Avg($\|y_i\|$)	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. $|Y|$ is the number of classes. Avg($|y_i|$) is the average number of classes per sample. Depth is the maximum level of label hierarchy.

Prompt Tuning Framework for HTC

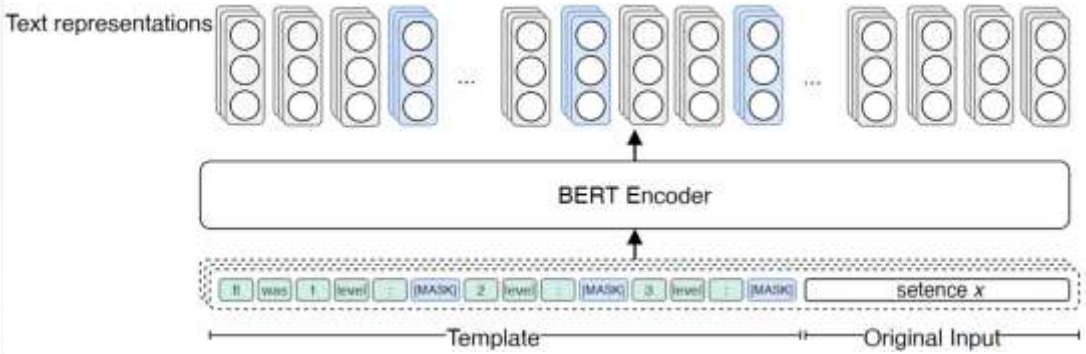
Hard Prompt:

[CLS] sentence x [SEP] The text is about [MASK] [SEP]

[CLS] It was 1 level: [MASK] 2 level: [MASK] [SEP] sentence x [SEP]

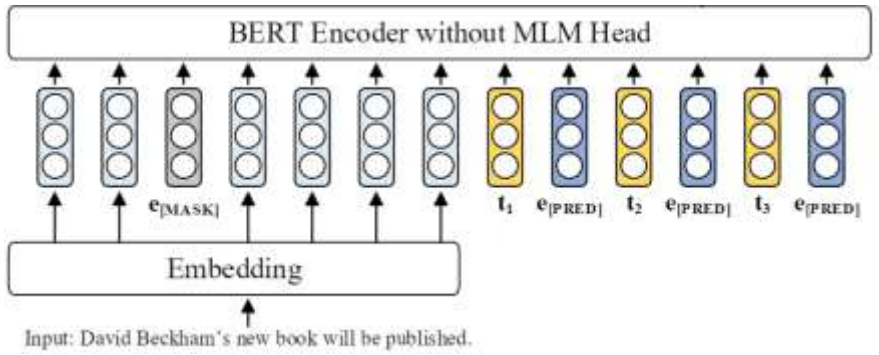
Soft Prompt:

[CLS] sentence x [SEP] [V1] [MASK] [V2] [MASK] … [V8] [MASK] [SEP]



HierVerb

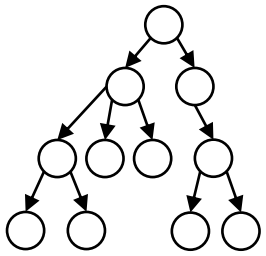
HPT



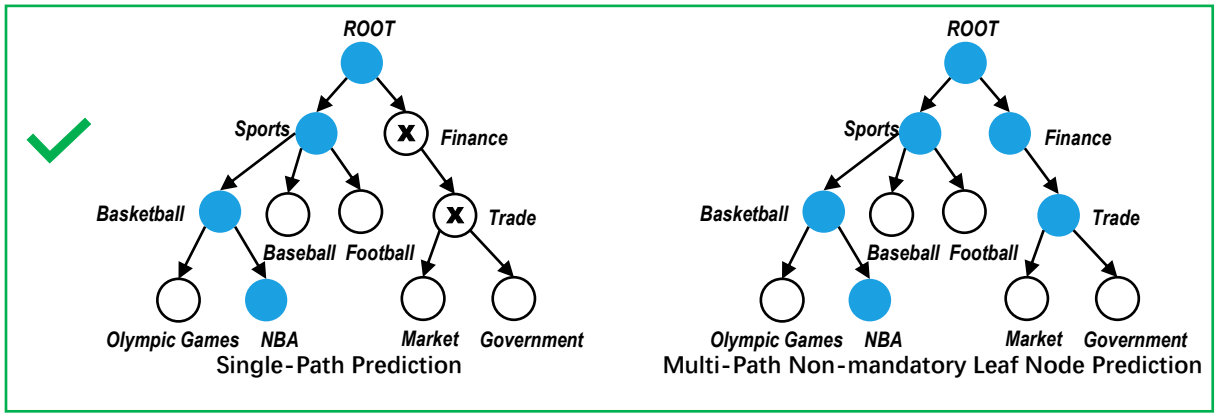
Challenges

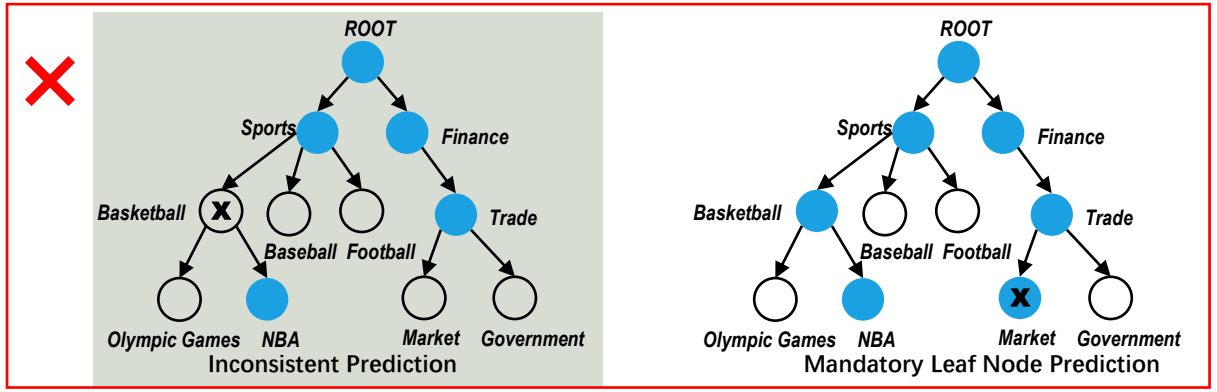
How to solve the hierarchical inconsistency problem in a simple and effective way

James Harden traded to Clippers.



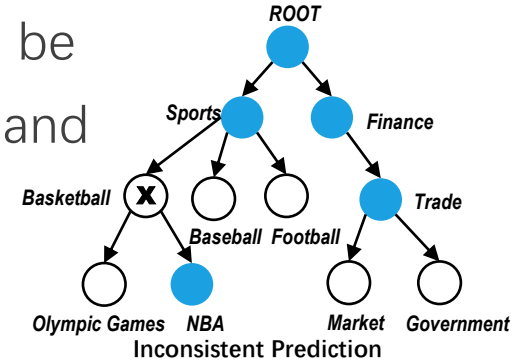
Label Hierarchy





Motivation

1. The solution to the problem of hierarchical consistency can be approached in terms of local relationships between parent and child nodes ;
2. The NER task can be analogized to the HTC task ;
3. NER tasks are more dependent on the relationship between neighboring tokens;



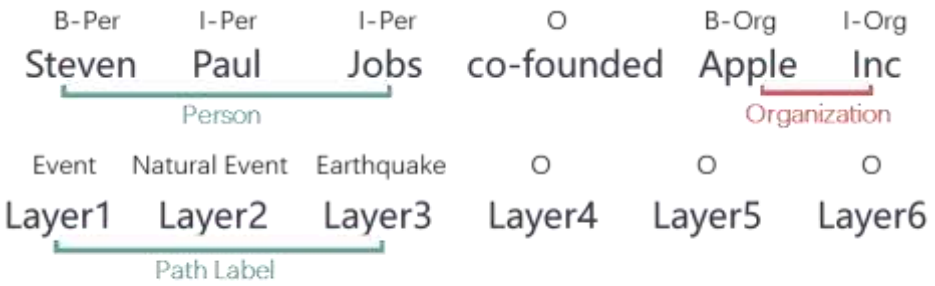
👇

Therefore, it is possible to migrate the methods of NER tasks to HTC tasks to strengthen the dependency between parent and child nodes.

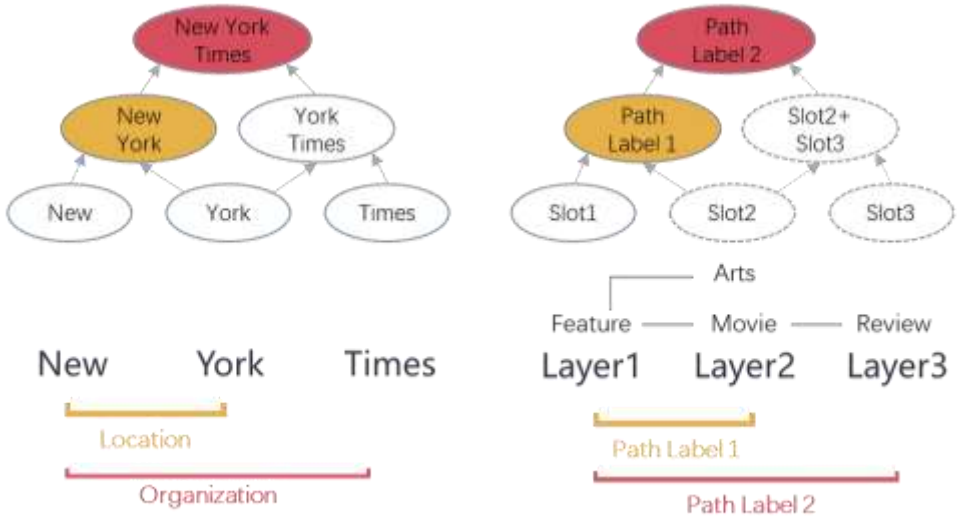
Motivation

Support 1:

[CLS] sentence **x** [SEP] [V1] [MASK] [V2] [MASK] ... [V8] [MASK] [SEP]



(a) Sequence labeling for flat NER and single-path HTC task



(b) Span-based classification for nested NER and multi-path HTC task

Motivation

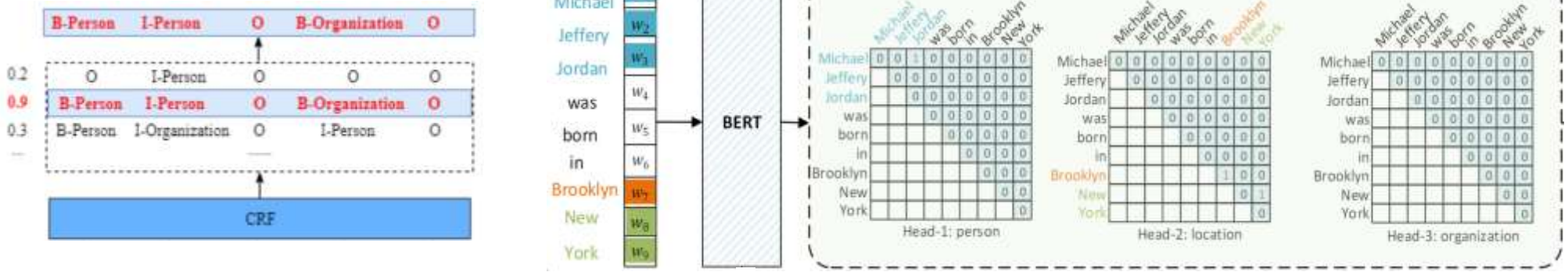
Support 2:

Louis David watched a game of the Chinese men's basketball team at the Yanyuan Garden of Peking University.

[B-PER，E-PER, O, O, O, O, O, B-ORG , I-ORG , I-ORG , E-ORG, O, O, B-LOC , E-LOC, O, B-ORG , I-ORG , I-ORG , E-ORG]

Michael Jeffery Jordan was born in Brooklyn New York

- Entity Index: B-Beginning, I-Intermediate, E-Ending, O-Others
- Entity Label: PER-Person, ORG-Organization, LOC-Location



Contributions

1. 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.
2. We employ Global Pointer and CRF designed for nested and flat entities to model both multipath and single-path HTC problems, **improving path consistency** in the results through a simple and effective way.
3. 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**.

Overall Framework

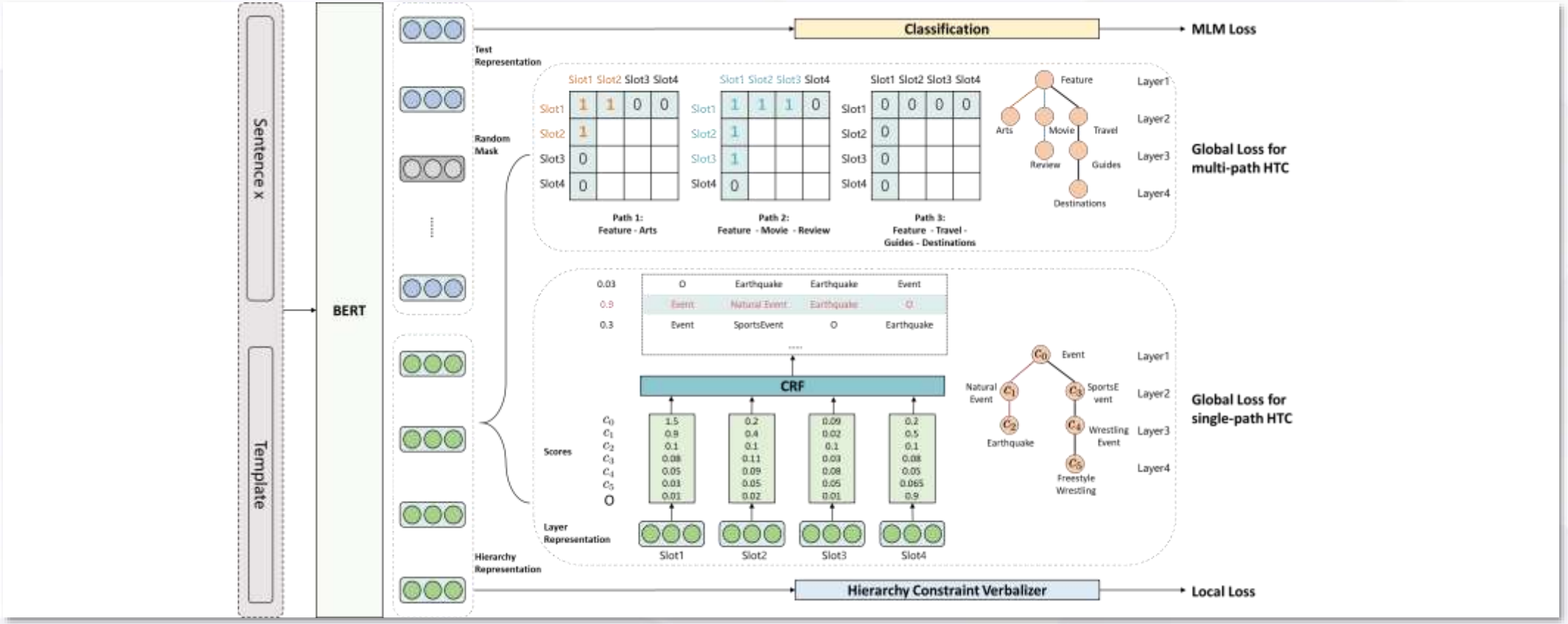
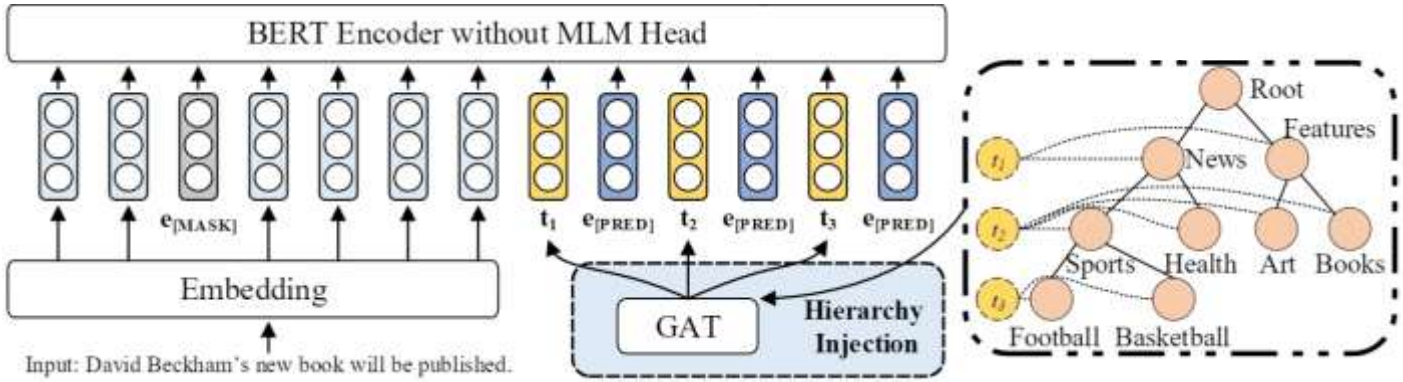


Figure 2: Model structure used in our method. We consider single-path HTC and multi-path HTC as flat and nested NER, respectively. For the multi-path task, we employ span-based Global Pointer method. In the case of single-path HTC, it is treated as a sequence labeling problem. For global-level information,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

Global Hierarchy-aware Structure



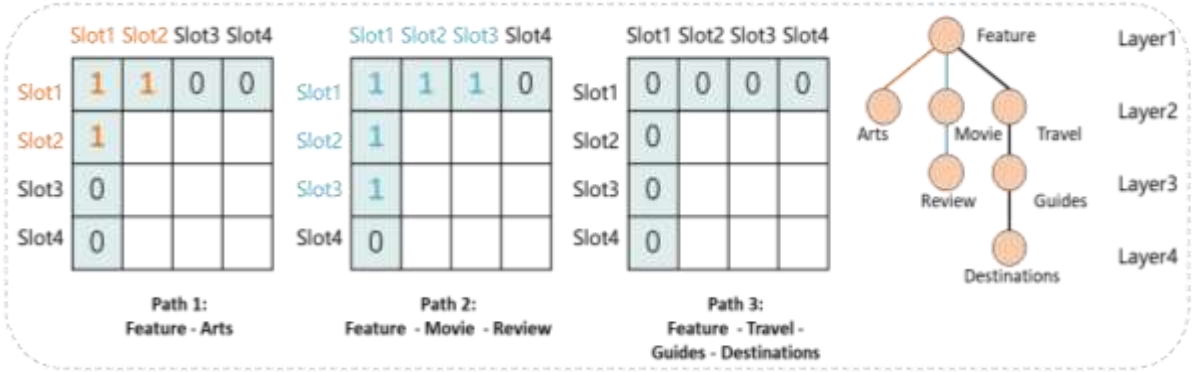
$$\mathbf{Emb} = [\mathbf{X}; \mathbf{T}] = \left[\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{t}_1, \mathbf{e}^1_{\text{hie}}, \ldots, \mathbf{t}_L, \mathbf{e}^L_{\text{hie}}\right]$$

$$\mathbf{G}_v^{(l+1)} = \text{ReLU}\left(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \frac{1}{c_v} \mathbf{W}^{(l)} \mathbf{G}_u^{(l)}\right)$$

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

$$\mathbf{H} = \text{BERT}(\mathbf{Emb}) = \left[\mathbf{h}_1, \ldots, \mathbf{h}_n, \mathbf{h}_{t_1}, \mathbf{h}^1_{\text{hie}}, \ldots, \mathbf{h}_{t_L}, \mathbf{h}^L_{\text{hie}}\right]$$

Local Hierarchy-aware Structure for Multi-path HTC



Global Loss for multi-path HTC

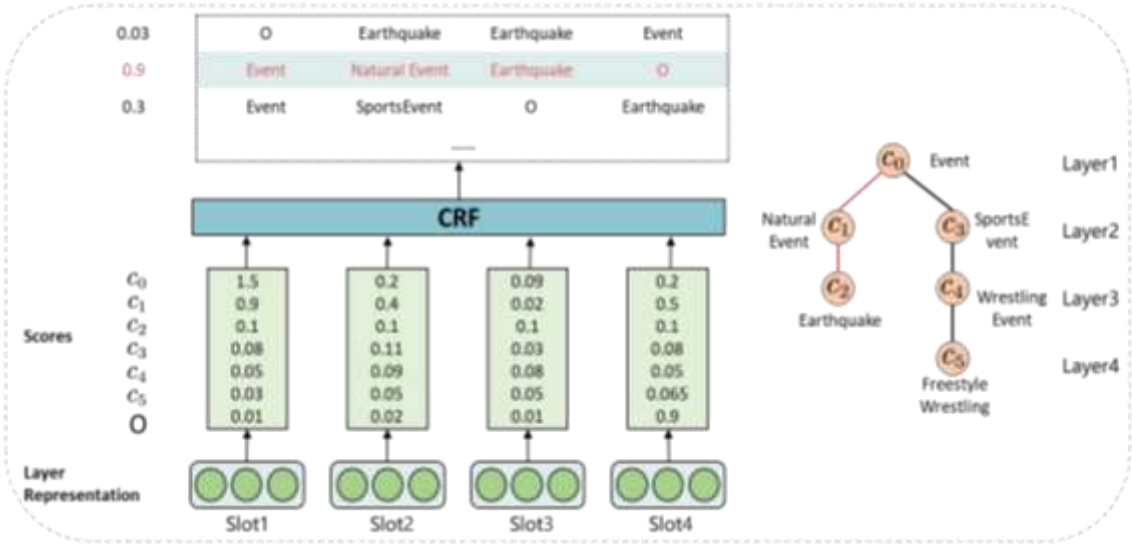
$$\mathbf{H} = \mathrm{BERT}(\mathbf{Emb}) = [\mathbf{h}_1, \ldots, \mathbf{h}_n, \mathbf{h}_{t_1}, \mathbf{h}^1_{\mathrm{hie}}, \ldots, \mathbf{h}_{t_L}, \mathbf{h}^L_{\mathrm{hie}}]$$

$$\mathbf{q}_{i,\alpha} = W_{q,\alpha}\mathbf{h}^i_{hie} + b_{q,\alpha} \qquad Q_\alpha = [\mathbf{q}_{1,\alpha}, \mathbf{q}_{2,\alpha}, \cdots, \mathbf{q}_{L,\alpha}]$$

$$\mathbf{k}_{j,\alpha} = W_{k,\alpha}\mathbf{h}^j_{hie} + b_{k,\alpha}, \qquad K_\alpha = [\mathbf{k}_{1,\alpha}, \mathbf{k}_{2,\alpha}, \cdots, \mathbf{k}_{L,\alpha}]$$

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

Local Hierarchy-aware Structure for Single-path HTC



Global Loss for single-path HTC

$$\mathbf{x}_{\text{hie}} = [x^1_{\text{hie}}, x^2_{\text{hie}}, \ldots, x^L_{\text{hie}}] \quad \mathbf{y} = \{y_1, y_2, \ldots, y_L\}$$

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

$$\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}) = \sum_{i=1}^{L} Em_{i,y_i} + \sum_{i=0}^{L} \text{Tr}_{y_i,y_{i+1}} \qquad \text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}) = \sum_{i=1}^{L} \boldsymbol{l}^i_{\text{hie}}[y_i] + \text{Tr}_{y_i,y_{i+1}}$$

Objective Function

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

$$\mathcal{L}_{MLM} = -\log \frac{e^{s_t}}{\sum_{i=1}^{V} e^{s_i}} = \log\left(1 + \sum_{i=1, i\neq t}^{V} e^{s_i - s_t}\right)$$

$$\mathcal{L}_{\text{Global}} = \sum_{m=1}^{L}\left(\log\left(1 + \sum_{i \in \mathcal{U}_m^{\text{neg}}} e^{s_i}\right) + \log\left(1 + \sum_{i \in \mathcal{U}_m^{\text{pos}}} e^{-s_i}\right)\right)$$

$$\mathcal{L}_{\text{Local}} = \sum_{\alpha \in P}\left(\log\left(1 + \sum_{(i,j) \in \mathcal{Q}_\alpha} e^{s_\alpha(i,j)}\right) + \log\left(1 + \sum_{(i,j) \in \Omega_\alpha} e^{-s_\alpha(i,j)}\right)\right)$$ multi-path

$$\mathcal{L}_{\text{Local}} = -\log(\Pr(\mathbf{y} \mid \mathbf{x}_{\text{hie}})) = \log\left(\sum_{\mathbf{y}' \in P} \exp^{\text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y}')}\right) - \text{Score}(\mathbf{x}_{\text{hie}}, \mathbf{y})$$ single-path

Experiments

Main Results and Consistency Experiments:

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.

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.

Other Experiments

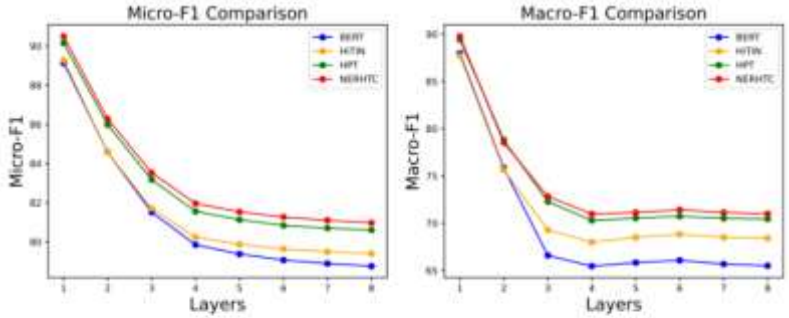


Figure 3: Peference study on label granularity of NYT based on layer increment.

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`.

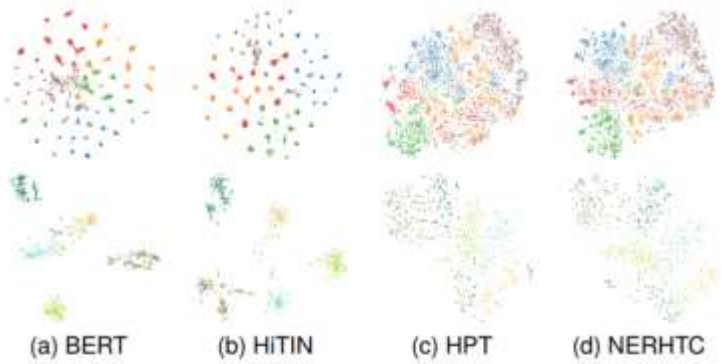


Figure 4: T-SNE visualization of the label representations on WOS. Images in the first row display feature clusters for the first-level labels, while the second row is for the sub-labels of ECE. Dots of the same color belong to the same category.

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

Thank You!