Global and Local Hierarchy-aware Contrastive Framework for Implicit Discourse Relation Recognition

Yuxin Jiang[1,2] Linhan Zhang[3] Wei Wang[1,2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]School of Computer Science and Engineering, The University of New South Wales
yjiangcm@connect.ust.hk, linhan.zhang@unsw.edu.au, weiwcs@ust.hk

ACL 2023

code: https://github.com/YJiangcm/GOLF_for_IDRR

Reported by Zicong Dou

Introduction

这吸引了人们的注意……

这只是导致该公司决定退出竞标的另一个风险因素。

That attracts attention . . . **it was just another one of the risk factors that led to the company's decision to withdraw from the bidding.**

Top-level sense: Contingency

Second-level sense: Cause

Implicit connective: but

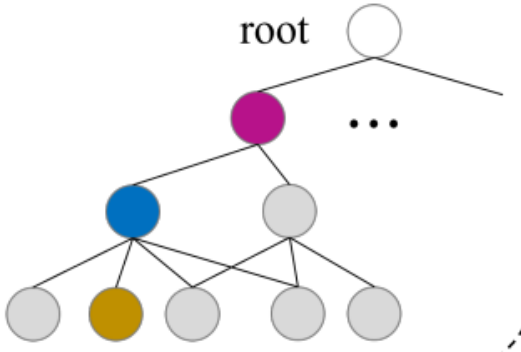


Figure 1: An IDRR instance in the PDTB 2.0 corpus (Prasad et al., 2008). Argument 1 is in italics, and argument 2 is in bold. The implicit connective is not present in the original discourse context but is assigned by annotators. All senses defined in PDTB are organized in a three-layer hierarchical structure (defined as *global hierarchy* in our paper), and the implicit connectives can be regarded as the most fine-grained senses.

(1) *Manufacturers' backlogs of unfilled orders rose 0.5% in September to $497.34 billion, helped by strength in the defense capital goods sector.* **Excluding these orders, backlogs declined 0.3%.**

Top: Comparison, Sec: Contrast, Conn: but

(2) *That attracts attention . . .* **it was just another one of the risk factors that led to the company's decision to withdraw from the bidding.**

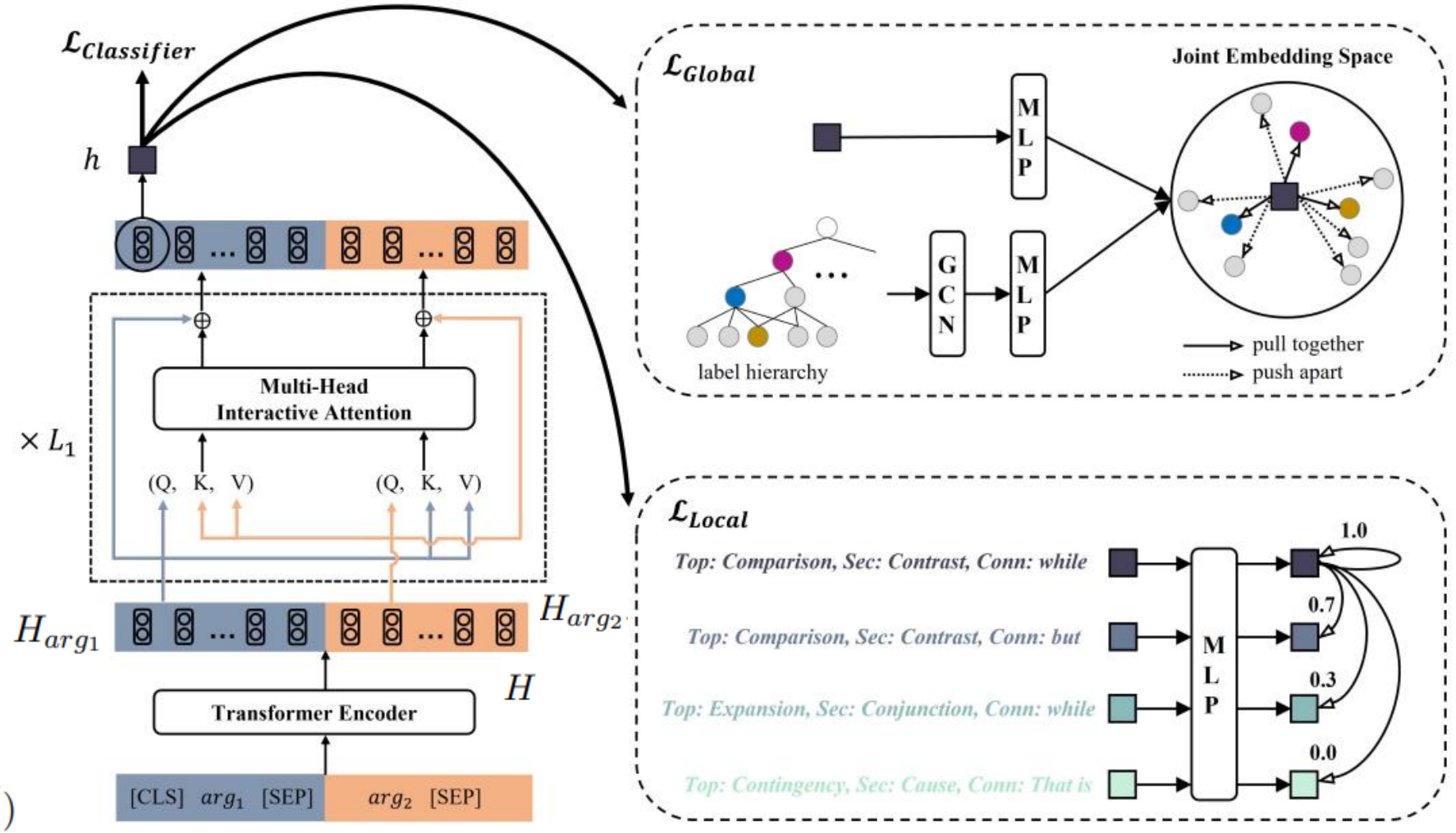
Top: Contingency, Sec: Cause, Conn: but

(3) *She offered Mrs. Yeargin a quiet resignation and thought she could help save her teaching certificate.* **Mrs. Yeargin declined.**

Top: Comparison, Sec: Contrast, Conn: however

Figure 2: Three instances from PDTB 2.0. The sense label sequence of each instance is defined as *local hierarchy* in our paper.

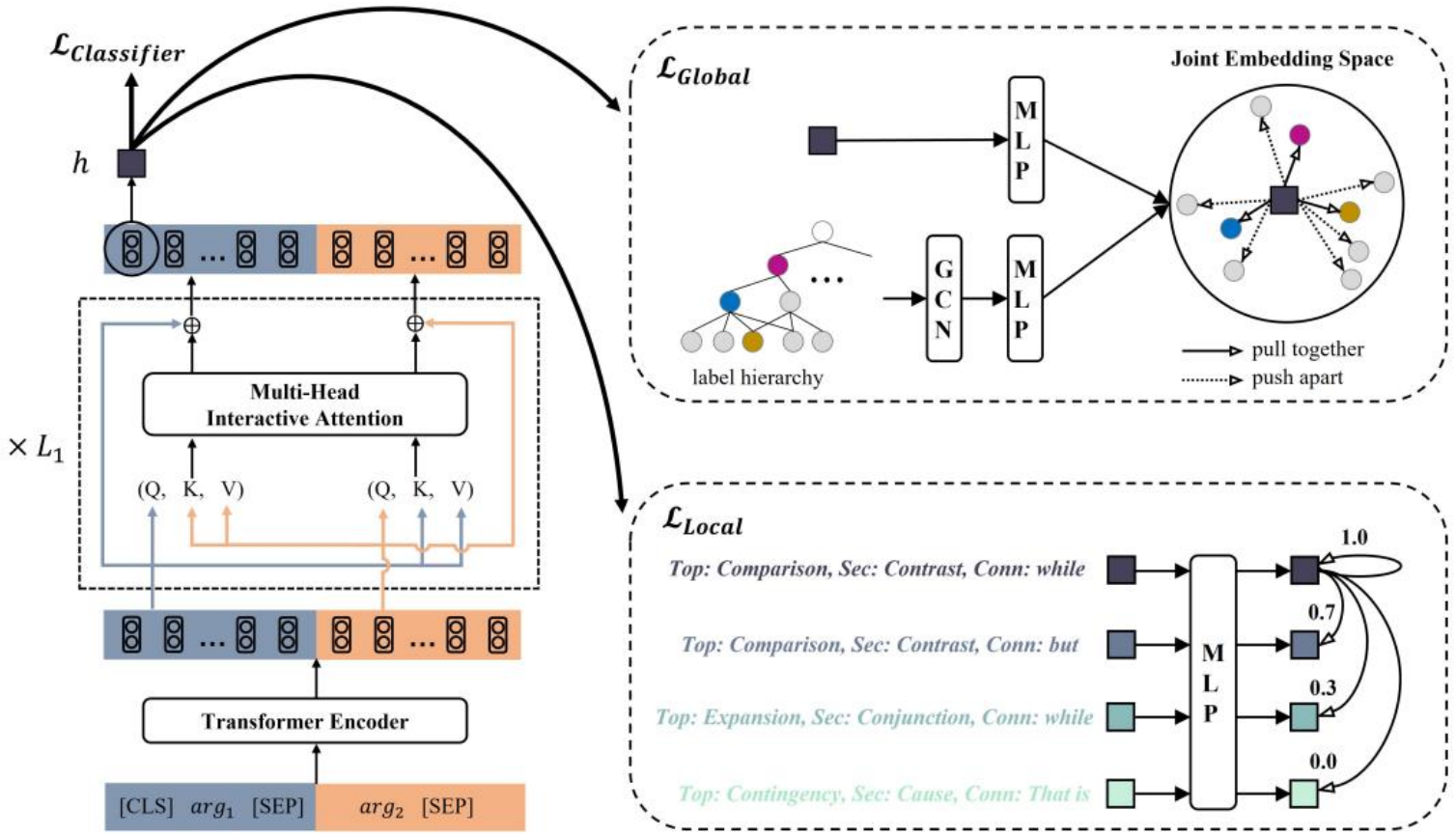
Method



$$S = (S^1, ..., S^m, ..., S^M)$$
$$x_i = (arg_1, arg_2)$$
$$y_i = (y_i^1, ..., y_i^m, ..., y_i^M)$$

Figure 3: The overall architecture of our framework. The squares are denoted as discourse relation representations. Among the local hierarchy-aware contrastive loss $\mathcal{L}_{Local}$, we use colored squares to denote discourse relation representations of various instances in a mini-batch and list their sense label sequences on the left. Besides, note that the numbers on the right are similarity scores between sense label sequences calculated by our scoring function.

Method



Staircase Classifier

the discourse relation representation h_i of an instance

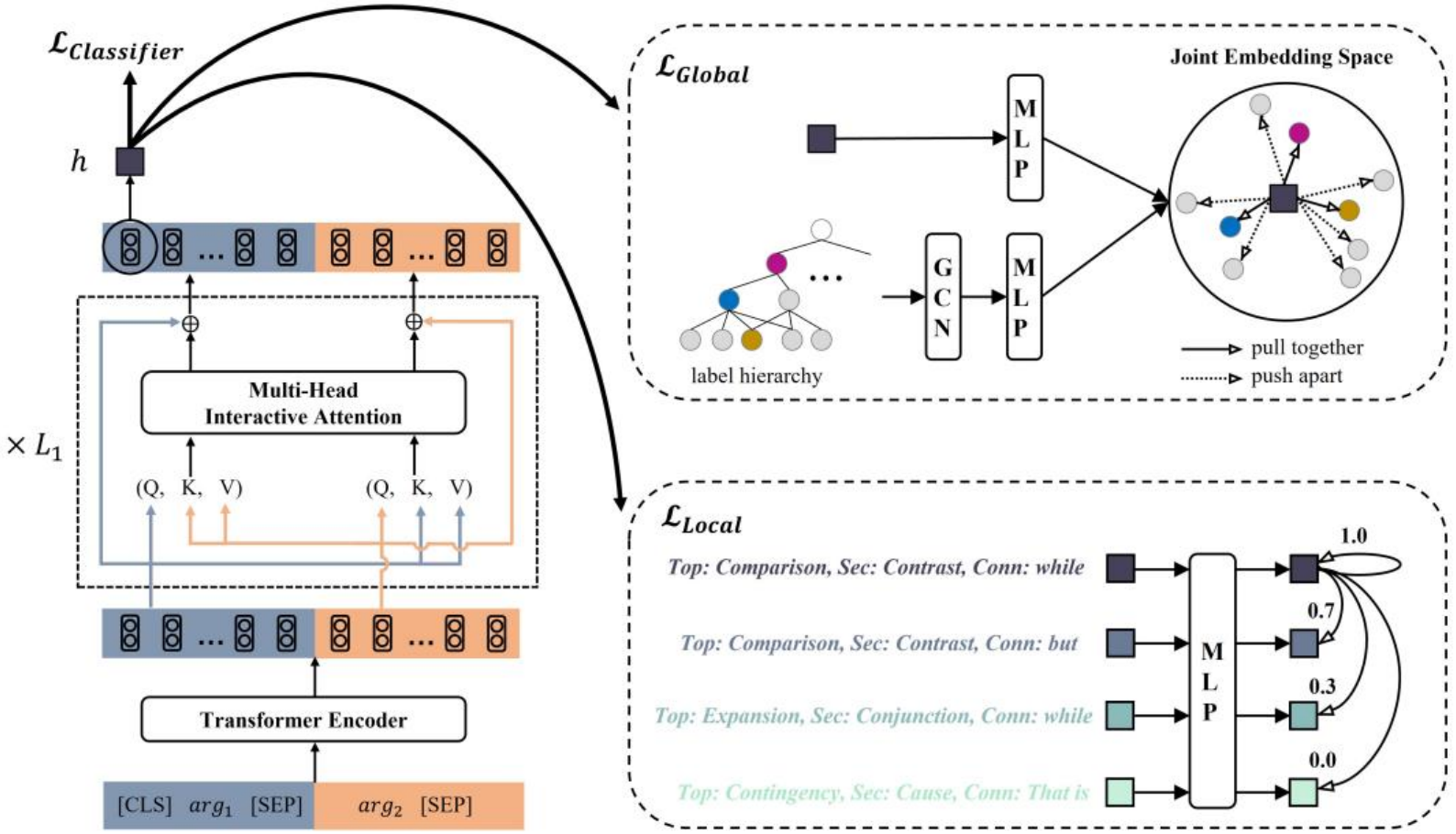
$$t_i^m = h_i W_1^m + t_i^{m-1} W_2^m + b^m \quad (1)$$

where $W_1^m \in \mathbb{R}^{d_h \times |S^m|}$, $W_2^m \in \mathbb{R}^{|S^{m-1}| \times |S^m|}$, $b^m \in \mathbb{R}^{|S^m|}$, $t_i^0 = \vec{0}$. Then the cross-entropy loss of the classifier is defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{|N|} \sum_{i \in N} \sum_{m=1}^{M} \mathbb{E}_{\vec{y}_i^m} [\text{LogSoftmax}(t_i^m)] \quad (2)$$

where $\vec{y}_i^m$ is the one-hot encoding of the ground-truth sense label y_i^m.

Method



Global Hierarchy-aware Contrastive Learning

Global Hierarchy Encoder

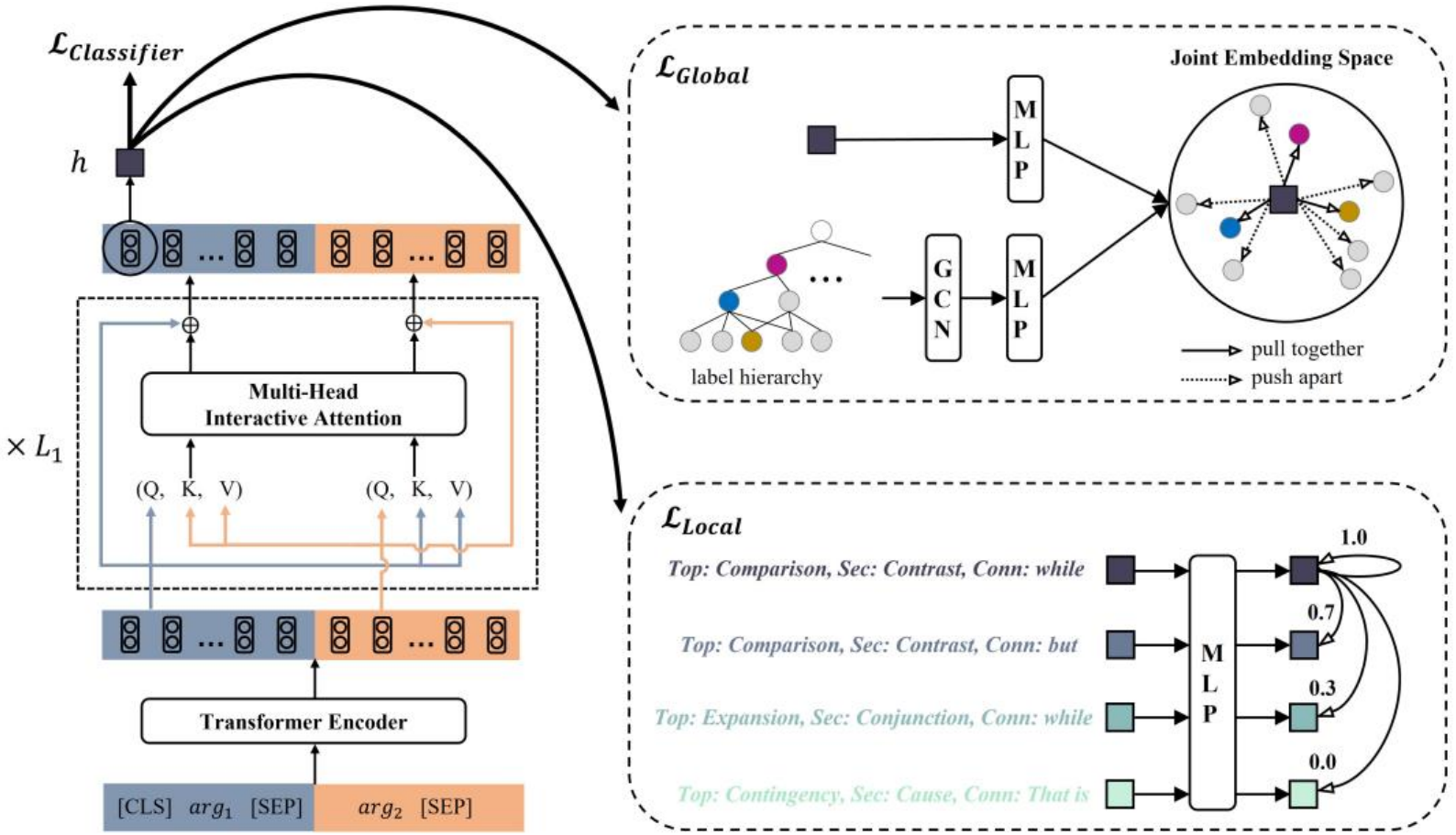
The adjacent matrix $A \in \mathbb{R}^{|S| \times |S|}$

$$A_{ij} = \begin{cases} 1, & if\ i = j; \\ 1, & if\ child(i) = j\ or\ child(j) = i; \\ 0, & otherwise. \end{cases} \quad (3)$$

where S is the set of all senses, $i, j \in S$, $child(i) = j$ means that sense j is the subclass of sense i. By setting the number layer of GCN as L_2, given the initial representation of sense i as $r_i^0 \in \mathbb{R}^{d_r}$, GCN updates the sense embeddings with the following layer-wise propagation rule:

$$r_i^l = ReLU(\sum_{j \in S} D_{ii}^{-\frac{1}{2}} A_{ij} D_{jj}^{-\frac{1}{2}} r_j^{l-1} W^l + b^l) \quad (4)$$

where $l \in [1, L_2]$, $W^l \in \mathbb{R}^{d_r \times d_r}$ and $b^l \in \mathbb{R}^{d_r}$ are learnable parameters at the l-th GCN layer, $D_{ii} = \sum_j A_{ij}$. Finally, we take the output $\{r_i^{L_2}\}_{i \in S}$ of the L_2-th layer as the sense embeddings and denote them as $\{r_i\}_{i \in S}$ for simplicity.



where N denotes a batch of training instances, y_i is the sense label sequence of instance x_i, $sim(\cdot)$ is the cosine similarity function, τ is a temperature hyperparameter.

Global Hierarchy-aware Contrastive Learning

Semantic Match in a Joint Embedding Space

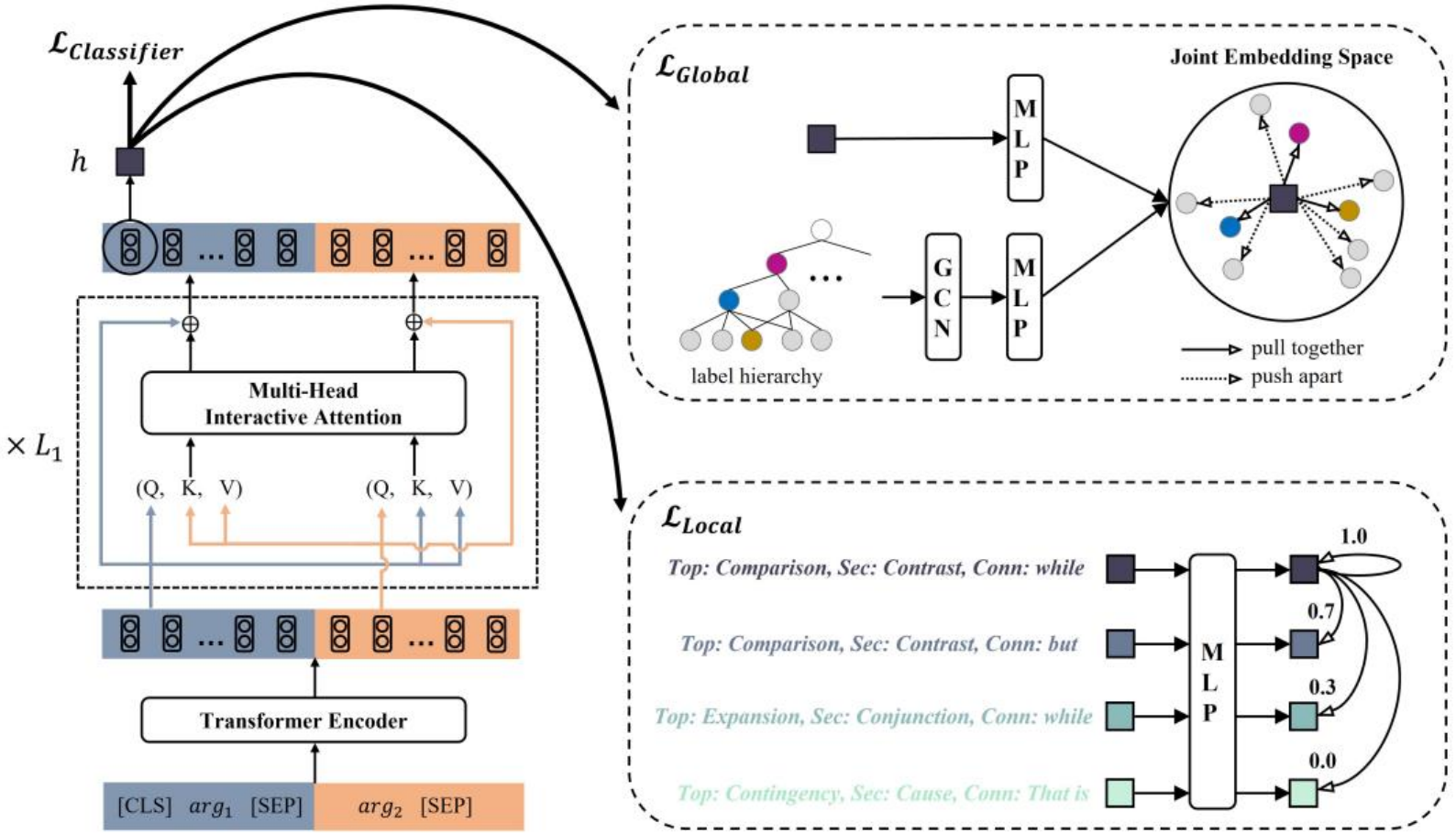
the discourse relation representation h_i of an instance x_i

the sense label embeddings $\{r_i\}_{i \in S}$

Multi-Layer Perception (MLP) Φ_1 and Φ_2.

$$\mathcal{L}_G = -\frac{1}{|N|} \sum_{i \in N} \sum_{j \in S} \mathbb{1}_{j \in y_i} \times \log \frac{\exp\left(sim\left(\Phi_1(h_i), \Phi_2(r_j)\right)/\tau\right)}{\sum_{j \in S} \exp\left(sim\left(\Phi_1(h_i), \Phi_2(r_j)\right)/\tau\right)} \quad (5)$$

Method



$$\mathcal{L}_L = -\frac{1}{|N|}\sum_{i\in N}\sum_{j\in N^+} Score(y_i, y_j) \times \log \frac{\exp\left(sim\left(\Phi_3(h_i), \Phi_3(h_j)\right)/\tau\right)}{\sum_{j\in N^+}\exp\left(sim\left(\Phi_3(h_i), \Phi_3(h_j)\right)/\tau\right)} \quad (8)$$

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \cdot \mathcal{L}_G + \lambda_2 \cdot \mathcal{L}_L \quad (9)$$

Local Hierarchy-aware Contrastive Learning

MLP layer Φ_3

$$\mathcal{L}_{L'} = -\frac{1}{|N|}\sum_{i\in N}\sum_{j\in N^+}\left(\prod_{m=1}^{M}\mathbb{1}_{y_i^m = y_j^m}\right) \times \log \frac{\exp\left(sim\left(\Phi_3(h_i), \Phi_3(h_j)\right)/\tau\right)}{\sum_{j\in N^+}\exp\left(sim\left(\Phi_3(h_i), \Phi_3(h_j)\right)/\tau\right)} \quad (6)$$

$$y_i = (y_i^1, ..., y_i^m, ..., y_i^M)$$
$$y_j = (y_j^1, ..., y_j^m, ..., y_j^M)$$

Top, Second, and Connective,

$P = \{\mathbb{T}, \mathbb{S}, \mathbb{C}, \mathbb{TS}, \mathbb{SC}, \mathbb{TSC}\}$ $K = 6$

$$Score(y_i, y_j) = \frac{1}{K}\sum_{k=1}^{K} Dice(P_i^k, P_j^k) \quad (7)$$

where $Dice(A, B) = (2|A \cap B|)/(|A| + |B|)$, P_i^k is the k-th sub-path label set of y_i.

$\frac{1}{6}(\frac{2\times1}{1+1} + \frac{2\times1}{1+1} + \frac{2\times0}{1+1} + \frac{2\times2}{2+2} + \frac{2\times1}{2+2} + \frac{2\times2}{3+3}) \approx 0.7.$

Experiments

Second-level Senses	Train	Dev	Test
Exp.Conjunction	2,814	258	200
Exp.Restatement	2,430	260	211
Exp.Instantiation	1,100	106	118
Exp.List	330	9	12
Exp.Alternative	150	10	9
Cont.Cause	3,234	281	269
Cont.Pragmatic cause	51	6	7
Comp.Contrast	1,569	166	128
Comp.Concession	181	15	17
Temp.Asynchronous	540	46	54
Temp.Synchrony	148	8	14
Total	12,547	1,165	1,039

Table 6: The data statistics of second-level senses in PDTB 2.0.

Second-level Senses	Train	Dev	Test
Exp.Conjunction	3,566	298	237
Exp.Level-of-detail	2,698	274	214
Exp.Instantiation	1,215	117	127
Exp.Manner	1,159	57	53
Exp.Substitution	405	32	31
Exp.Equivalence	256	25	30
Cont.Cause	4,280	423	388
Cont.Purpose	688	66	59
Cont.Cause+Belief	140	13	14
Cont.Condition	138	17	14
Comp.Concession	1,159	105	97
Comp.Contrast	813	87	62
Temp.Asynchronous	1,025	103	105
Temp.Synchronous	331	24	35
Total	17,873	1,641	1,466

Table 7: The data statistics of second-level senses in PDTB 3.0.

Experiments

Model	Embedding	Top-level		Second-level		Connective	
		F_1	Acc	F_1	Acc	F_1	Acc
PDTB 2.0							
NNMA (Liu and Li, 2016)	GloVe	46.29	57.57	-	-	-	-
KANN (Guo et al., 2020)	GloVe	47.90	57.25	-	-	-	-
PDRR (Dai and Huang, 2018)	word2vec	48.82	57.44	-	-	-	-
IDRR-Con (Shi and Demberg, 2019)	word2vec	46.40	61.42	-	47.83	-	-
IDRR-C&E (Dai and Huang, 2019)	ELMo	52.89	59.66	33.41	48.23	-	-
MTL-MLoss (Nguyen et al., 2019)	ELMo	53.00	-	-	49.95	-	-
HierMTN-CRF (Wu et al., 2020)	BERT	55.72	65.26	33.91	53.34	10.37	30.00
BERT-FT (Kishimoto et al., 2020)	BERT	58.48	65.26	-	54.32	-	-
RoBERTa (Fine-tuning)	RoBERTa	62.96	69.98	40.34	59.87	10.06	31.45
BMGF-RoBERTa (Liu et al., 2020)	RoBERTa	63.39	69.06	-	58.13	-	-
LDSGM (Wu et al., 2022)	RoBERTa	63.73	71.18	40.49	60.33	10.68	32.20
ChatGPT (Chan et al., 2023a)	-	36.11	44.18	16.20	24.54	-	-
GOLF (base)	RoBERTa	65.76	72.52	41.74	61.16	11.79	32.85
GOLF (large)	RoBERTa	**69.60**	**74.67**	**47.91**	**63.91**	**14.59**	**42.35**
PDTB 3.0							
MANF (Xiang et al., 2022a)	BERT	56.63	64.04	-	-	-	-
RoBERTa (Fine-tuning)	RoBERTa	68.31	71.59	50.63	60.14	14.72	39.43
BMGF-RoBERTa (Liu et al., 2020)	RoBERTa	63.39	69.06	-	58.13	-	-
LDSGM (Wu et al., 2022)	RoBERTa	68.73	73.18	53.49	61.33	17.68	40.20
ConnPrompt (Xiang et al., 2022b)	RoBERTa	69.51	73.84	-	-	-	-
GOLF (base)	RoBERTa	70.88	75.03	55.30	63.57	19.21	42.54
GOLF (large)	RoBERTa	**74.21**	**76.39**	**60.11**	**66.42**	**20.66**	**45.12**

Table 1: Model comparison of multi-class classification on PDTB 2.0 and PDTB 3.0 in terms of macro-averaged F1 (%) and accuracy (%).

Experiments

Model	Exp. (53%)	Cont. (27%)	Comp. (14%)	Temp. (3%)
BMGF (Liu et al., 2020)	77.66	60.98	59.44	50.26
LDSGM (Wu et al., 2022)	78.47	64.37	61.66	50.88
GOLF (base)	79.41	62.90	67.71	54.55
GOLF (large)	**80.96**	**66.54**	**69.47**	**61.40**

Table 2: Label-wise F1 scores (%) for the top-level senses of PDTB 2.0. The proportion of each sense is listed below its name.

Second-level Senses	BMGF	LDSGM	GOLF (base)	GOLF (large)
Exp.Restatement (20%)	53.83	58.06	**59.84**	59.03
Exp.Conjunction (19%)	60.17	57.91	60.28	**61.54**
Exp.Instantiation (12%)	67.96	72.60	75.36	**77.98**
Exp.Alternative (1%)	60.00	63.46	**63.49**	61.54
Exp.List (1%)	0.00	8.98	27.78	**43.48**
Cont.Cause (26%)	59.60	64.36	65.35	**65.98**
Cont.Pragmatic (1%)	0.00	0.00	0.00	0.00
Comp.Contrast (12%)	59.75	63.52	61.95	**67.57**
Comp.Concession (2%)	0.00	0.00	0.00	**11.11**
Temp.Asynchronous (5%)	56.18	56.47	63.82	**65.49**
Temp.Synchrony (1%)	0.00	0.00	0.00	**13.33**

Table 3: Label-wise F1 scores (%) for the second-level senses of PDTB 2.0. The proportion of each sense is listed behind its name.

Experiments

Top-level Senses	**GOLF** (base)	**GOLF** (large)
Exp (47%)	80.01	**80.50**
Cont (32%)	74.54	**74.83**
Comp (11%)	64.67	**71.59**
Temp (10%)	64.80	**70.92**

Table 8: Label-wise F1 scores (%) for the top-level senses of PDTB 3.0. The proportion of each sense is listed behind its name.

Second-level Senses	**GOLF** (base)	**GOLF** (large)
Exp.Conjunction (16%)	**64.09**	63.69
Exp.Level-of-detail (15%)	52.60	**59.29**
Exp.Instantiation (9%)	72.53	**73.77**
Exp.Manner (4%)	**63.53**	62.61
Exp.Substitution (2%)	66.67	**72.22**
Exp.Equivalence (2%)	**25.39**	24.00
Cont.Cause (26%)	69.47	**72.49**
Cont.Purpose (4%)	71.60	**72.73**
Cont.Cause+Belief (1%)	0.00	0.00
Cont.Condition (1%)	66.67	**92.31**
Comp.Concession (7%)	59.09	**63.37**
Comp.Contrast (4%)	43.33	**60.27**
Temp.Asynchronous (7%)	68.79	**77.55**
Temp.Synchronous (2%)	41.00	**42.27**

Table 9: Label-wise F1 scores (%) for the second-level senses of PDTB 3.0. The proportion of each sense is listed behind its name.

Experiments

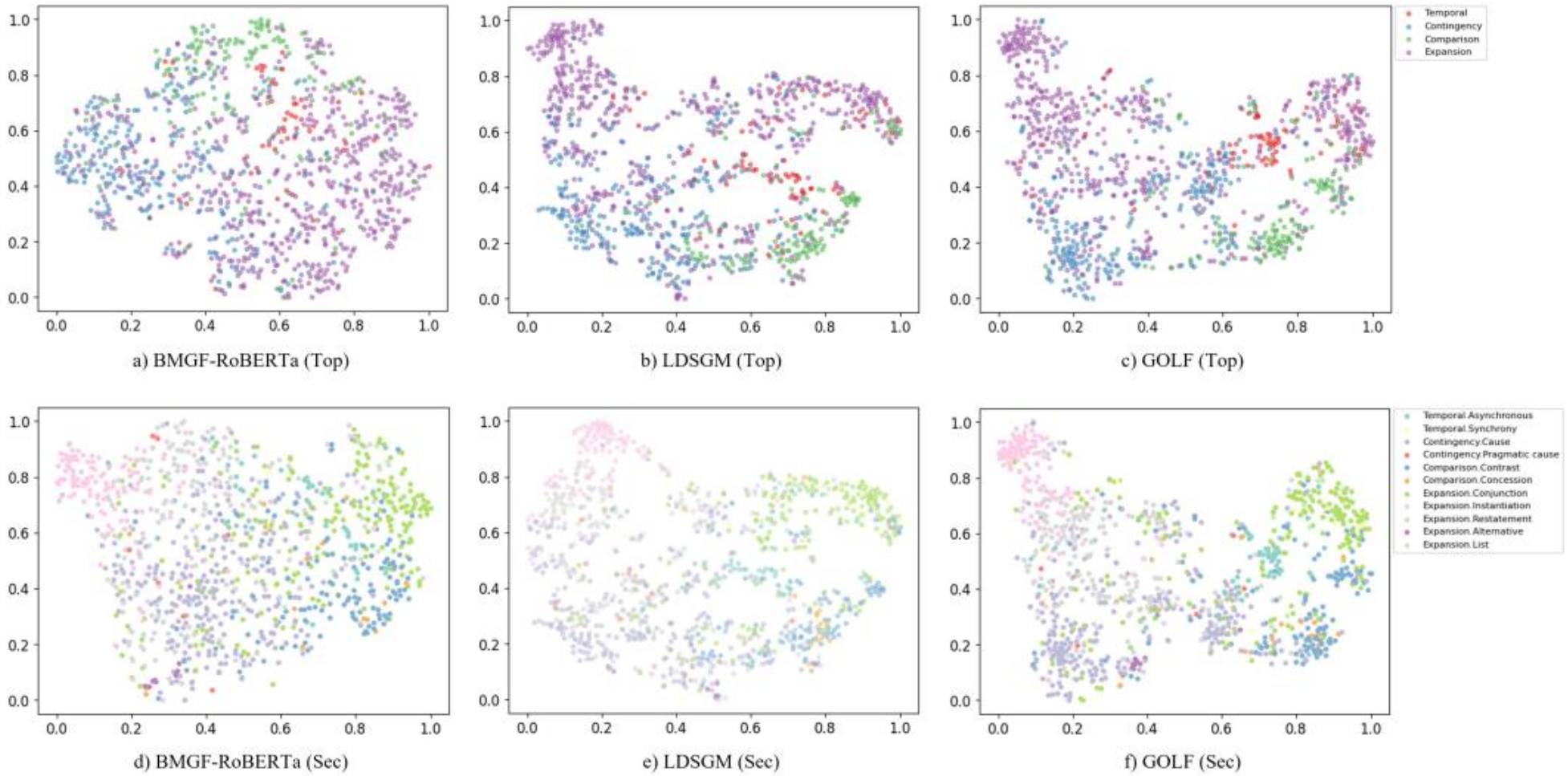
Model	Top-level		Second-level		Connective		Top-Sec	Top-Sec-Conn
	F_1	Acc	F_1	Acc	F_1	Acc		
GOLF	**65.76**	**72.52**	**41.74**	**61.16**	**11.79**	**32.85**	**59.65**	**27.55**
-*w/o* MHIA	64.97	71.85	41.07	60.52	10.80	31.69	58.52	26.18
-*w/o* staircase	65.43	72.25	41.12	60.81	10.81	31.40	58.43	26.08
-*w/o* MHIA and staircase	64.77	71.98	40.99	60.10	10.76	31.65	58.49	26.22
-*w/o* $\mathcal{L}_G$	65.37	71.61	40.78	60.40	11.56	32.73	59.01	26.86
-*w/o* $\mathcal{L}_L$	64.34	71.32	40.24	60.42	10.76	31.88	58.69	26.37
-*w/o* $\mathcal{L}_G$ and $\mathcal{L}_L$	63.85	71.04	39.98	59.92	10.72	30.47	58.23	25.89
-*r.p.* $\mathcal{L}_L$ with $\mathcal{L}_{L'}$	64.58	71.56	41.20	61.07	11.43	32.55	59.24	27.05

Table 4: Ablation study on PDTB 2.0 considering the accuracy and F1 of each level as well as consistencies between hierarchies. "*w/o*" stands for "without"; "*r.p.*" stands for "replace"; "MHIA" stands for the Multi-Head Interactive Attention; $\mathcal{L}_G$ stands for the Global Hierarchy-aware Contrastive loss; $\mathcal{L}_L$ stands for the Local Hierarchy-aware Contrastive loss.

Experiments

Model	Top-Sec	Top-Sec-Conn
PDTB 2.0		
HierMTN-CRF	46.29	19.15
BMGF-RoBERTa	47.06	21.37
LDSGM	58.61	26.85
GOLF (base)	59.65	27.55
GOLF (large)	**61.79**	**36.00**
PDTB 3.0		
HierMTN-CRF	50.19	27.82
BMGF-RoBERTa	52.33	29.16
LDSGM	60.32	34.57
GOLF (base)	61.31	36.97
GOLF (large)	**64.86**	**38.26**

Table 5: Comparison with current state-of-the-art models on the consistency among multi-level sense predictions.



a) BMGF-RoBERTa (Top) b) LDSGM (Top) c) GOLF (Top)

d) BMGF-RoBERTa (Sec) e) LDSGM (Sec) f) GOLF (Sec)

Figure 4: t-SNE visualization of discourse relation representations for the top-level and second-level senses on PDTB 2.0 test set.

Experiments

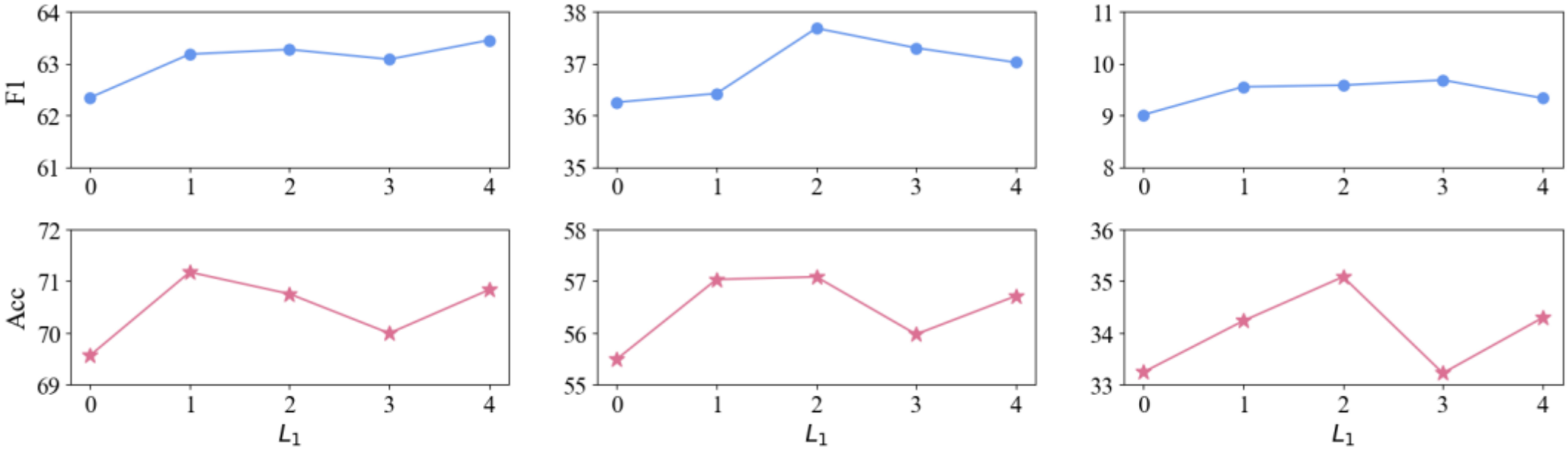


Figure 5: Effects of the number layer L_1 of MHIA on the development set.

Experiments

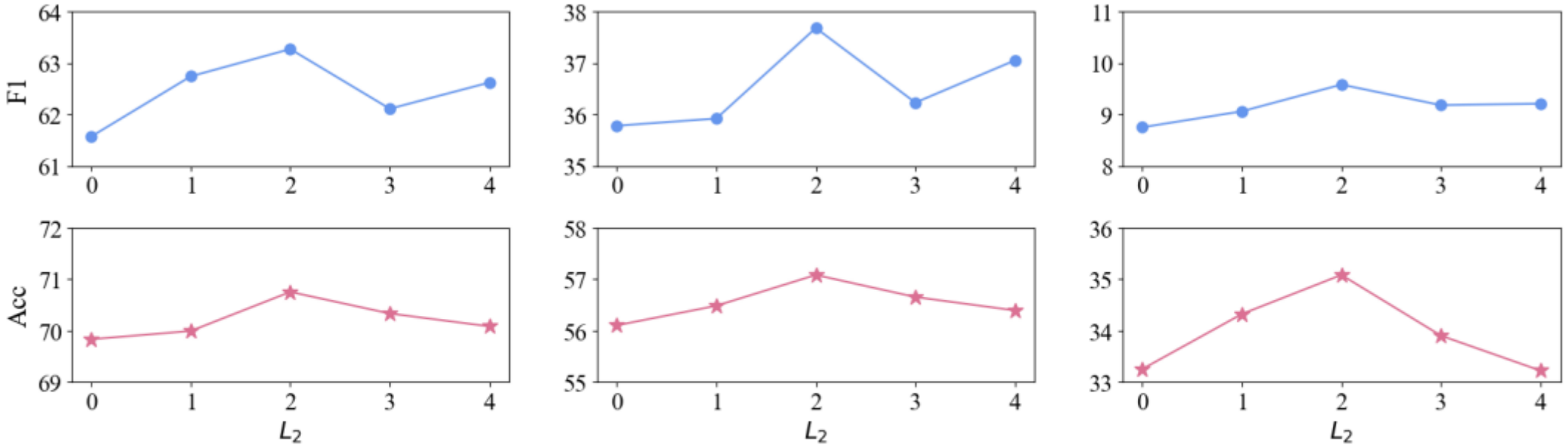


Figure 6: Effects of the number layer L_2 of GCN on the development set.

Experiments

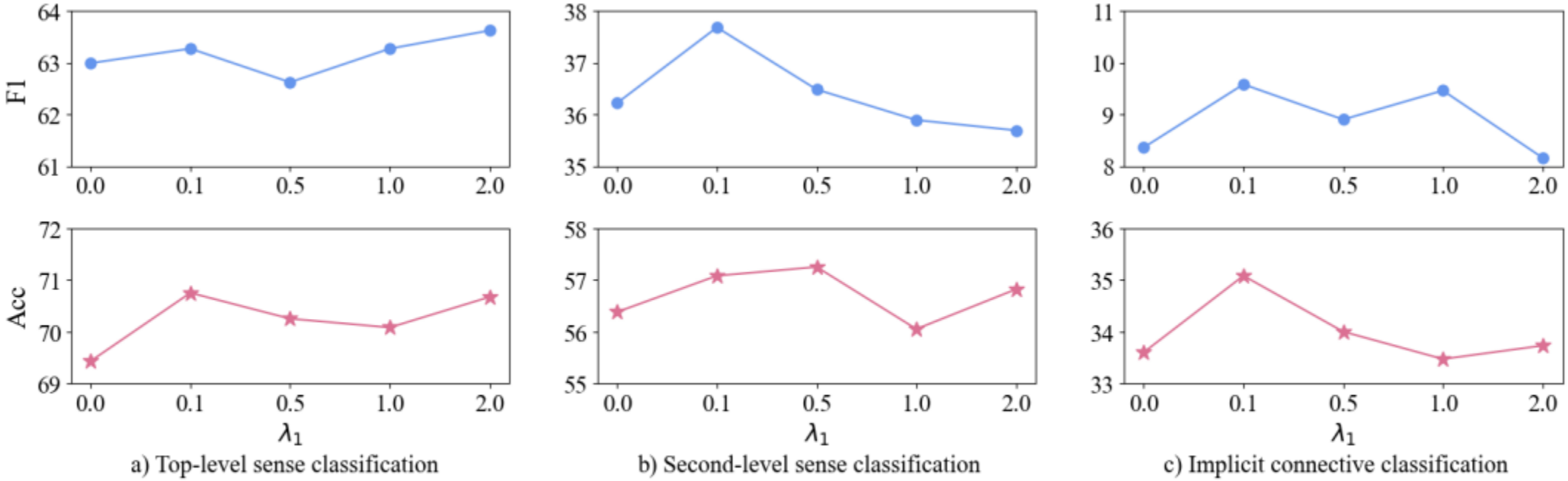
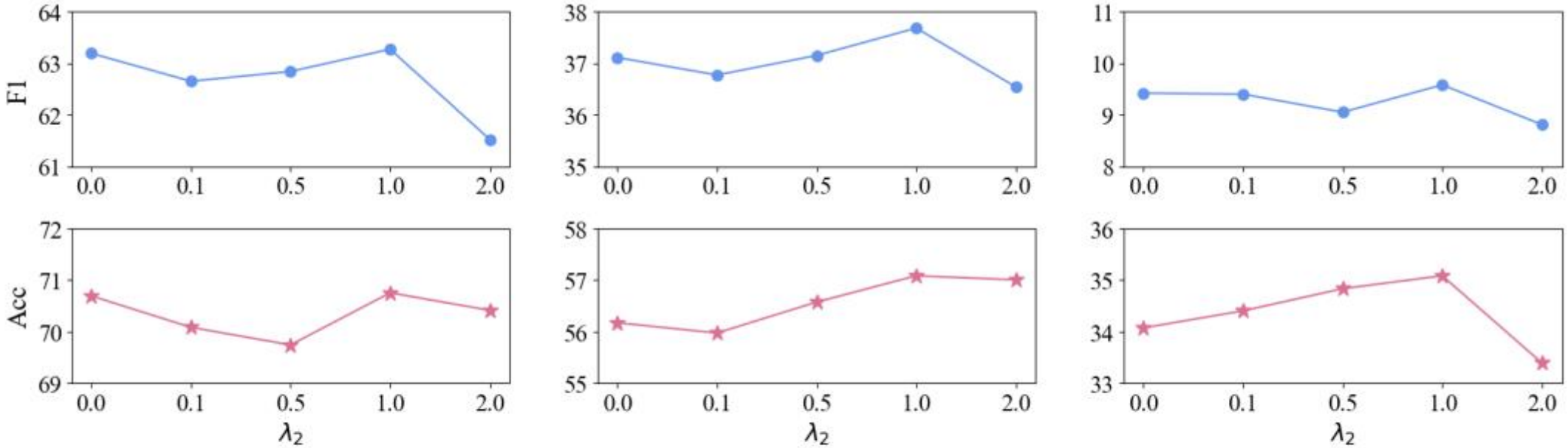


Figure 7: Effects of the coefficient λ_1 of the global hierarchy-aware contrastive loss on the development set.

Experiments



a) Top-level sense classification

b) Second-level sense classification

c) Implicit connective classification

Figure 8: Effects of the coefficient λ_2 of the local hierarchy-aware contrastive loss on the development set.

Experiments

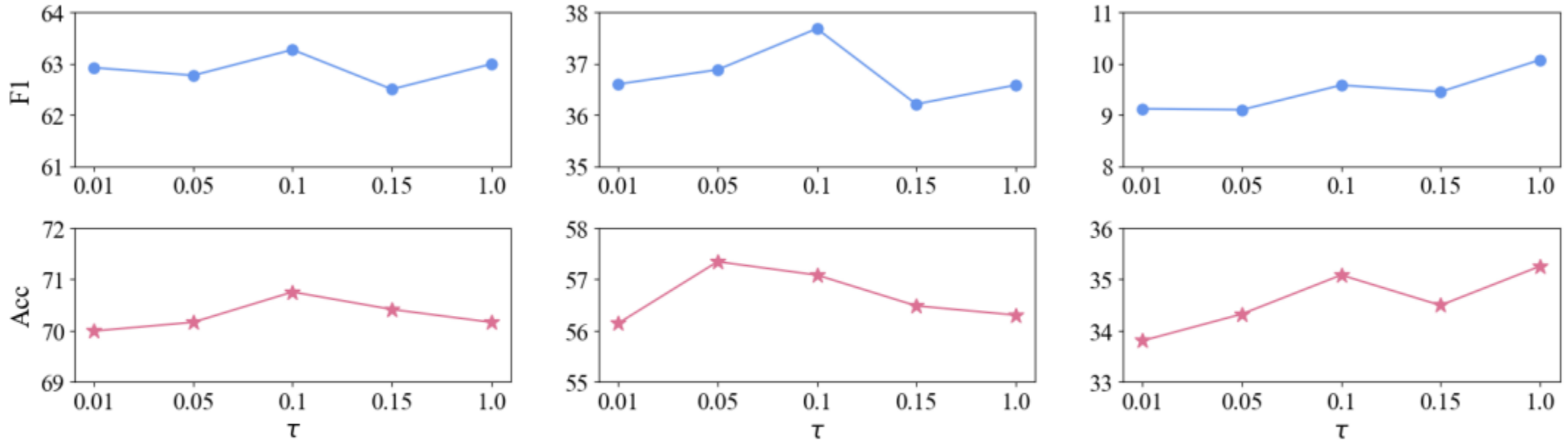


Figure 9: Effects of the temperature τ in contrastive learning on the development set.