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

Yuxin Jiang[1,2] **Linhan Zhang**[3] **Wei Wang**[1,2,4]
[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
[4]Guangzhou Municipal Key Laboratory of Materials Informatics,
The Hong Kong University of Science and Technology (Guangzhou)
`yjiangcm@connect.ust.hk`, `linhan.zhang@unsw.edu.au`, `weiwcs@ust.hk`

Abstract

Due to the absence of explicit connectives, implicit discourse relation recognition (IDRR) remains a challenging task in discourse analysis. The critical step for IDRR is to learn high-quality discourse relation representations between two arguments. Recent methods tend to integrate the whole hierarchical information of senses into discourse relation representations for multi-level sense recognition. Nevertheless, they insufficiently incorporate the static hierarchical structure containing all senses (defined as *global hierarchy*), and ignore the hierarchical sense label sequence corresponding to each instance (defined as *local hierarchy*). For the purpose of sufficiently exploiting global and local hierarchies of senses to learn better discourse relation representations, we propose a novel **Gl**O**bal and **L**ocal Hierarchy-aware Contrastive **F**ramework (GOLF), to model two kinds of hierarchies with the aid of *multi-task learning* and *contrastive learning*. Experimental results on PDTB 2.0 and PDTB 3.0 datasets demonstrate that our method remarkably outperforms current state-of-the-art models at all hierarchical levels. [1]

1 Introduction

Implicit discourse relation recognition (IDRR) aims to identify logical relations (named senses) between a pair of text segments (named arguments) without an explicit connective (e.g., `however`, `because`) in the raw text. As a fundamental task in discourse analysis, IDRR has benefitted a wide range of Natural Language Processing (NLP) applications such as question answering (Liakata et al., 2013), summarization (Cohan et al., 2018), information extraction (Tang et al., 2021), etc.

The critical step for IDRR is to learn high-quality discourse relation representations between two arguments. Early methods are dedicated to manually designing shallow linguistic features (Pitler et al., 2009; Park and Cardie, 2012) or constructing dense representations relying on word embeddings (Liu and Li, 2016; Dai and Huang, 2018; Liu et al., 2020). Despite their successes, they train multiple models to predict multi-level senses independently, while ignoring that the sense annotation of IDRR follows a hierarchical structure (as illustrated in Figure 1). To solve this issue, some researchers propose global hierarchy-aware models to exploit the prior probability of label dependencies based on Conditional Random Field (CRF) (Wu et al., 2020) or the sequence generation model (Wu et al., 2022).

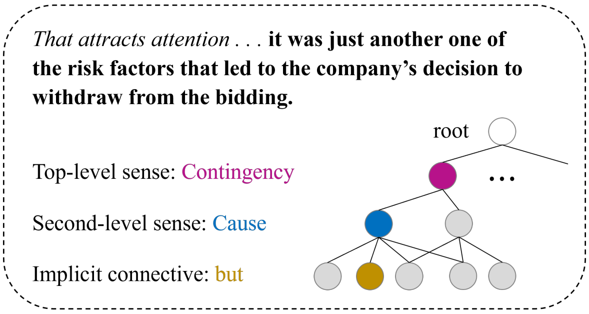

That attracts attention . . . **it was just another one of the risk factors that led to the company's decision to withdraw from the bidding.**

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.

However, existing hierarchy-aware methods still have two limitations. *Firstly*, though they exploit the fact that there are complex dependencies among senses and such information should be encoded into discourse relation representations, their manners of encoding the holistic hierarchical graph of senses may not be sufficient, since they fail to strengthen the correlation between the discourse relation representation and its associated sense labels, which is

[1]Our code is publicly available at `https://github.com/YJiangcm/GOLF_for_IDRR`

Findings of the Association for Computational Linguistics: ACL 2023, pages 8048–8064
July 9-14, 2023 ©2023 Association for Computational Linguistics

(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.

highly useful for classification (Chen et al., 2020a). *Secondly*, they only consider the graph of the entire label hierarchy and ignore the benefit of the label sequence corresponding to each instance. As shown in Figure 2, the label sequences of Instances (1) and (2) differ at both the top and second levels, while the label sequences of Instances (1) and (3) only differ at the most fine-grained level. The similarity between label sequences provides valuable information for regularizing discourse relation representations, e.g., by ensuring that the distance between representations of Instance (1) and (2) is farther than the distance between representations of Instance (1) and (3). Under such an observation, we categorize the sense hierarchy into global and local hierarchies to fully utilize the hierarchical information in IDRR. We define *global hierarchy* as the entire hierarchical structure containing all senses, while *local hierarchy* is defined as a hierarchical sense label sequence corresponding to each input instance. Therefore, global hierarchy is static and irrelevant to input instances, while local hierarchy is dynamic and pertinent to input instances.

Built on these motivations, we raise our research question: *How to sufficiently incorporate global and local hierarchies to learn better discourse relation representations?* To this end, we propose a novel **G**l**O**bal and **L**ocal Hierarchy-aware Contrastive **F**ramework (GOLF), to inject additional information into the learned relation representation through additional tasks that are aware of the global and local hierarchies, respectively. This is achieved via the joint use of *multi-task learning* and *contrastive learning*. The key idea of contrastive learning is to narrow the distance between two semantically similar representations, meanwhile, pushing away representations of dissimilar pairs (Chen et al., 2020b; Gao et al., 2021). It has achieved extraordinary successes in representation learning (He et al., 2020). Finally, our multi-task learning framework consists of classification tasks and two additional contrastive learning tasks. The global hierarchy-aware contrastive learning task explicitly matches textual semantics and label semantics in a text-label joint embedding space, which refines the discourse relation representations to be semantically similar to the target label representations while semantically far away from the incorrect label representations. In the local hierarchy-aware contrastive learning task, we propose a novel scoring function to measure the similarity among sense label sequences. Then the similarity is utilized to guide the distance between discourse relation representations.

The main contributions of this paper are threefold:

- We propose a novel global and local hierarchy-aware contrastive framework for IDRR, which sufficiently incorporates global and local hierarchies to learn better discourse relation representations.
- To our best knowledge, our work is the first attempt to meticulously adapt contrastive learning to IDRR considering the global and local hierarchies of senses.
- Comprehensive experiments and thorough analysis demonstrate that our approach delivers state-of-the-art performance on PDTB 2.0 and PDTB 3.0 datasets at all hierarchical levels, and more consistent predictions on multi-level senses.

2 Related Work

2.1 Implicit Discourse Relation Recognition

Early studies resort to manually-designed features to classify implicit discourse relations into four top-level senses (Pitler et al., 2009; Park and Cardie, 2012). With the rapid development of deep learning, many methods explore the direction of building deep neural networks based on static word embeddings. Typical works include shallow CNN (Zhang et al., 2015), LSTM with Multi-Level Attention (Liu and Li, 2016), knowledge-augmented LSTM

(Dai and Huang, 2018, 2019; Guo et al., 2020), etc. These works aim to learn better semantic representations of arguments as well as capture the semantic interaction between them. More recently, contextualized representations learned from large pre-trained language models (PLMs) and prompting (Schick and Schütze, 2021) have substantially improved the performance of IDRR. More fined-grained levels of senses have been explored by (Liu et al., 2020; Long and Webber, 2022; Chan et al., 2023b). Besides, researchers such as (Wu et al., 2020, 2022) utilize the dependence between hierarchically structured sense labels to predict multi-level senses simultaneously. However, these methods may be insufficient to exploit the global and local hierarchies for discourse relation representations.

2.2 Contrastive Learning

Contrastive learning is initially proposed in Computer Vision (CV) as a weak-supervised representation learning method, aiming to pull semantically close samples together and push apart dissimilar samples (He et al., 2020; Chen et al., 2020b). In NLP, contrastive learning has also achieved extraordinary successes in various tasks including semantic textual similarity (STS) (Gao et al., 2021; Shou et al., 2022; Jiang et al., 2022), information retrieval (IR) (Hong et al., 2022), relation extraction (RE) (Chen et al., 2021), etc. Though intuitively supervised contrastive learning could be applied to IDRR through constructing positive pairs according to the annotated sense labels, it ignores the hierarchical structure of senses. This paper is the first work to meticulously adapt contrastive learning to IDRR considering the global and local hierarchies of senses.

3 Problem Definition

Given M hierarchical levels of defined senses $S = (S^1, ..., S^m, ..., S^M)$, where S^m is the set of senses at the m-th hierarchical level, and a sample input consisting of two text spans, or $x_i = (arg_1, arg_2)$, our model aims to output a sequence of sense $y_i = (y_i^1, ..., y_i^m, ..., y_i^M)$, where $y_i^m \in S^m$.

4 Methodology

Figure 3 illustrates the overall architecture of our multi-task learning framework. Beginning at the left part of Figure 3, we utilize a Discourse Relation Encoder to capture the interaction between two input arguments and map them into a discourse relation representation h. After that, the discourse relation representation h is fed into a Staircase Classifier to perform classification at three hierarchical levels dependently. While training, we will use two additional tasks, the global hierarchy-aware contrastive loss $\mathcal{L}_{Global}$ (in the upper right part of Figure 3) and the local hierarchy-aware contrastive loss $\mathcal{L}_{Local}$ (in the lower right part of Figure 3) as additional regularization to refine the discourse relation representation h. During inference, we only use the Discourse Relation Encoder and the Staircase Classifier for classification and *discard* the Global and Local Hierarchy-aware Contrastive Learning modules. Detailed descriptions of our framework are given below.

4.1 Discourse Relation Encoder

Given an instance $x_i = (arg_1, arg_2)$, we concatenate the two arguments and formulate them as a sequence with special tokens: `[CLS]` arg_1 `[SEP]` arg_2 `[SEP]`, where `[CLS]` and `[SEP]` denote the beginning and the end of sentences, respectively. Then we feed the sequence through a Transformer (Vaswani et al., 2017) encoder to acquire contextualized token representations H. Previous works (Liu and Li, 2016; Liu et al., 2020) indicate that deep interactions between two arguments play an important role in IDRR. To this end, we propose a Multi-Head Interactive Attention (MHIA) module to facilitate bilateral multi-perspective matching between arg_1 and arg_2. As shown in the left part of Figure 3, we separate H into H_{arg_1} and H_{arg_2}, denoting as the contextualized representations of arg_1 and arg_2. Then MHIA reuses the Multi-Head Attention (MHA) in Transformer, but the difference is that we take H_{arg_1} as *Query*, H_{arg_2} as *Key* and *Value* and vice versa. The intuition behind MHIA is to simulate human's transposition thinking process: respectively considering each other's focus from the standpoint of arg_1 and arg_2. Note that the MHIA module may be stacked for L_1 layers. Finally, we use the representation of `[CLS]` in the last layer as the discourse relation representation and denote it as h for simplicity.

4.2 Staircase Classifier

Given the discourse relation representation h_i of an instance, we propose a "staircase" classifier inspired by (Abbe et al., 2021) to output the label logits t_i^m at each hierarchical level $m \in [1, M]$ in

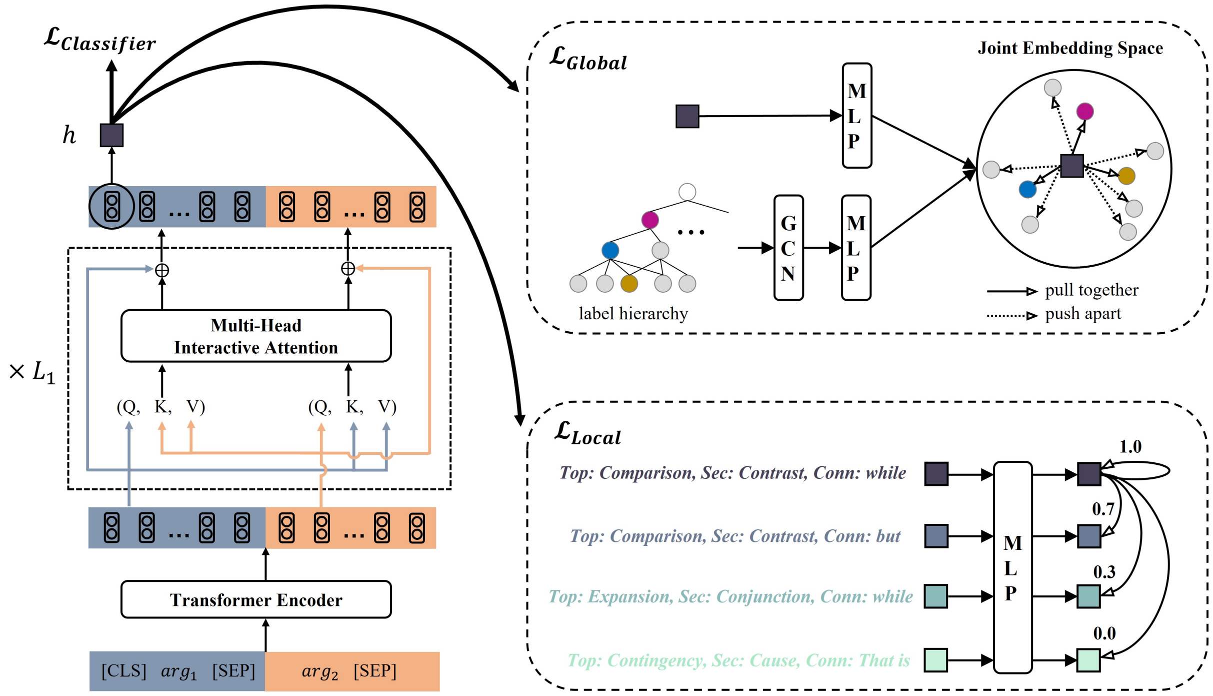

$\mathcal{L}_{Classifier}$
h
$\times L_1$
Multi-Head Interactive Attention
(Q, K, V)
(Q, K, V)
Transformer Encoder
[CLS] arg_1 [SEP]
arg_2 [SEP]
$\mathcal{L}_{Global}$
Joint Embedding Space
MLP
GCN
MLP
label hierarchy
pull together
push apart
$\mathcal{L}_{Local}$
Top: Comparison, Sec: Contrast, Conn: while
Top: Comparison, Sec: Contrast, Conn: but
Top: Expansion, Sec: Conjunction, Conn: while
Top: Contingency, Sec: Cause, Conn: That is
MLP
1.0
0.7
0.3
0.0

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.

a top-down manner, where the higher-level logits are used to guide the logits at the current level:

$$t_i^m = h_i W_1^m + t_i^{m-1} W_2^m + b^m \tag{1}$$

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:

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

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

4.3 Global Hierarchy-aware Contrastive Learning

The Global Hierarchy-aware Contrastive Learning module first exploits a Global Hierarchy Encoder to encode global hierarchy into sense label embeddings. Then, it matches the discourse relation representation of an input instance with its corresponding sense label embeddings in a joint embedding space based on contrastive learning.

4.3.1 Global Hierarchy Encoder

To encode label hierarchy in a global view, we regard the hierarchical structure of senses as an undirected graph, where each sense corresponds to a graph node. Then we adopt a graph convolutional network (GCN) (Welling and Kipf, 2016) to induce node embeddings for each sense based on properties of their neighborhoods. The adjacent matrix $A \in \mathbb{R}^{|S| \times |S|}$ is defined as follows:

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

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:

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

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.

4.3.2 Semantic Match in a Joint Embedding Space

In this part, we match textual semantics and label semantics in a text-label joint embedding space where correlations between text and labels are exploited, as depicted in the upper right part of Figure 3. We first project the discourse relation representation h_i of an instance x_i and the sense label embeddings $\{r_i\}_{i \in S}$ into a common latent space by two different Multi-Layer Perception (MLP) Φ_1 and Φ_2. Then, we apply a contrastive learning loss to capture text-label matching relationships, by regularizing the discourse relation representation to be semantically similar to the target label representations and semantically far away from the incorrect label representations:

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

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. By minimizing the global hierarchy-aware contrastive learning loss, the distribution of discourse relation representations is refined to be similar to the label distribution.

Here we would like to highlight the key differences between our model and LDSGM (Wu et al., 2022), since we both utilize a GCN to acquire label representations. Firstly, We use a different approach to capture the associations between the acquired label representations and the input text. In (Wu et al., 2022), the associations are *implicitly* captured using the usual attention mechanism. In contrast, our model *explicitly* learns them by refining the distribution of discourse relation representations to match the label distribution using contrastive learning. Secondly, our work introduces a novel aspect that has been overlooked by earlier studies including (Wu et al., 2022): the utilization of local hierarchy information, which enables our model to better differentiate between similar discourse relations and achieve further improvements.

4.4 Local Hierarchy-aware Contrastive Learning

Following (Gao et al., 2021), we duplicate a batch of training instances N as N^+ and feed N as well as N^+ through our Discourse Relation Encoder E with diverse dropout augmentations to obtain $2|N|$ discourse relation representations. Then we apply an MLP layer Φ_3 over the representations, which is shown to be beneficial for contrastive learning (Chen et al., 2020b).

To incorporate local hierarchy into discourse relation representations, it is tempting to directly apply supervised contrastive learning (Gunel et al., 2021) which requires positive pairs to have identical senses at each hierarchical level $m \in [1, M]$:

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

However, Equation (6) ignores the more subtle semantic structures of the local hierarchy, since it only admits positive examples as having *identical*, no account for examples with highly similar annotations. To illustrate, consider Instances (1) and (3) in Figure 2, where their sense label sequences only differ at the most fine-grained level. However, they are regarded as a negative pair in Equation (6), rather than a "relatively" positive pair. The standard of selecting positive pairs is too strict in Equation (6), thus may result in semantically similar representations being pulled away. To loosen this restriction, we regard all instance pairs as positive pairs but assign the degree of positive, by using a novel scoring function to calculate the similarity among label sequences $y_i = (y_i^1, ..., y_i^m, ..., y_i^M)$ and $y_j = (y_j^1, ..., y_j^m, ..., y_j^M)$.

In our case, there exist three hierarchical levels including Top, Second, and Connective, and we use $\mathbb{T}$, $\mathbb{S}$, and $\mathbb{C}$ to denote them. Consequently, there are in total $K = 6$ sub-paths in the hierarchies, i.e., $P = \{\mathbb{T}, \mathbb{S}, \mathbb{C}, \mathbb{TS}, \mathbb{SC}, \mathbb{TSC}\}$. Then we calculate the Dice similarity coefficient for each sub-path among the hierarchical levels and take the average as the similarity score between y_i and y_j, which is formulated below:

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

where $Dice(A, B) = (2|A \cap B|)/(|A| + |B|)$, P_i^k is the k-th sub-path label set of y_i. Taking Instances (1) and (3) in Figure 2 as examples, their label sequences are *Top: Comparison, Sec: Contrast,* *Conn: but* and *Top: Comparison, Sec: Contrast, Conn: however*, respectively. Then the similarity score would be $\frac{1}{6}(\frac{2\times1}{1+1}+\frac{2\times1}{1+1}+\frac{2\times0}{1+1}+\frac{2\times2}{2+2}+\frac{2\times1}{2+2}+\frac{2\times2}{3+3})\approx 0.7$.

Finally, our local hierarchy-aware contrastive loss utilizes the similarity scores to guide the distance between discourse relation representations:

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

Compared with Equation (6), Equation (8) considers more subtle semantic structures of the local hierarchy for selecting positive pairs. It increases the relevance of representations for all similarly labeled instances and only pushes away instances with entirely different local hierarchies. Thus, the local hierarchical information is sufficiently incorporated into discourse relation representations.

The overall training goal is the combination of the classification loss, the global hierarchy-aware contrastive loss, and the local hierarchy-aware contrastive loss:

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

where λ_1 and λ_2 are coefficients for the global and local hierarchy-aware contrastive loss, respectively. We set them as 0.1 and 1.0 while training, according to hyperparameter search (in Appendix C).

5 Experiments

5.1 Dataset

The Penn Discourse Treebank 2.0 (PDTB 2.0) PDTB 2.0 (Prasad et al., 2008) is a large-scale English corpus annotated with information on discourse structure and semantics. PDTB 2.0 has three levels of senses, i.e., classes, types, and sub-types. Since only part of PDTB instances is annotated with third-level senses, we take the top-level and second-level senses into consideration and regard the implicit connectives as third-level senses. There are 4 top-level senses including Temporal (Temp), Contingency (Cont), Comparison (Comp), and Expansion (Expa). Further, there exist 16 second-level senses, but we only consider 11 major second-level implicit types following previous works (Liu et al., 2020; Wu et al., 2022). For the connective classification, we consider all 102 connectives defined in PDTB 2.0.

The Penn Discourse Treebank 3.0 (PDTB 3.0) PDTB 3.0 (Webber et al., 2019) is the updated version of PDTB 2.0, which includes an additional 13K annotations and corrects some inconsistencies in PDTB 2.0. Following the preprocess of PDTB 2.0, we consider 4 top-level senses, 14 majority second-level senses, and all 186 connectives defined in PDTB 3.0.

Appendix A shows the detailed statistics of the PDTB corpora. We follow early works (Ji and Eisenstein, 2015; Liu et al., 2020; Wu et al., 2022) using Sections 2-20 of the corpus for training, Sections 0-1 for validation, and Sections 21-22 for testing. In PDTB 2.0 and PDTB 3.0, there are around 1% data samples with multiple annotated senses. Following (Qin et al., 2016), we treat them as separate instances during training for avoiding ambiguity. At test time, a prediction matching one of the gold types is regarded as the correct answer.

5.2 Baselines

To validate the effectiveness of our method, we contrast it with the most advanced techniques currently available. As past research generally assessed one dataset (either PDTB 2.0 or PDTB 3.0), we utilize distinct baselines for each. Due to PDTB 3.0's recent release in 2019, there are fewer baselines available for it compared to PDTB 2.0.

Baselines for PDTB 2.0

- **NNMA** (Liu and Li, 2016): a neural network with multiple levels of attention.
- **KANN** (Guo et al., 2020): a knowledge-enhanced attentive neural network.
- **PDRR** (Dai and Huang, 2018): a paragraph-level neural network that models inter-dependencies between discourse units as well as discourse relation continuity and patterns.
- **IDRR-Con** (Shi and Demberg, 2019): a neural model that leverages the inserted connectives to learn better argument representations.
- **IDRR-C&E** (Dai and Huang, 2019): a neural model leveraging external event knowledge and coreference relations.
- **MTL-MLoss** (Nguyen et al., 2019): a neural model which predicts the labels and connectives simultaneously.

- **HierMTN-CRF** (Wu et al., 2020): a hierarchical multi-task neural network with a conditional random field layer.

- **BERT-FT** (Kishimoto et al., 2020): a model applying three additional training tasks.

- **RoBERTa (Fine-tuning)**: a RoBERTa-based model fine-tuned on three sense levels separately.

- **BMGF-RoBERTa** (Liu et al., 2020): a RoBERTa-based model with bilateral multi-perspective matching and global information fusion.

- **LDSGM** (Wu et al., 2022): a label dependence-aware sequence generation model.

- **ChatGPT** (Chan et al., 2023a): a ChatGPT-based method equipped with an in-context learning prompt template.

Baselines for PDTB 3.0

- **MANF** (Xiang et al., 2022a): a multi-attentive neural fusion model to encode and fuse both semantic connection and linguistic evidence.

- **RoBERTa (Fine-tuning)**: a RoBERTa-based model fine-tuned on three sense levels separately.

- **BMGF-RoBERTa** (Liu et al., 2020): we reproduce the model on PDTB 3.0.

- **LDSGM** (Wu et al., 2022): we reproduce the model on PDTB 3.0.

- **ConnPrompt** (Xiang et al., 2022b): a PLM-based model using a connective-cloze Prompt to transform the IDRR task as a connective-cloze prediction task.

5.3 Implementation Details

We implement our model based on Huggingface's transformers (Wolf et al., 2020) and use the pre-trained RoBERTa (Liu et al., 2019) (base or large version) as our Transformer encoder. The layer number of MHIA and GCN are both set to 2. We set temperature τ in contrastive learning as 0.1. We set Φ_1, Φ_2, Φ_3 as a simple MLP with one hidden layer and *tanh* activation function, which enables the gradient to be easily backpropagated to the encoder. The node embeddings of senses with the dimension 100 are randomly initialized by *kaiming_normal* (He et al., 2015). To avoid overfitting, we apply dropout with a rate of 0.1 after each GCN layer. We adopt AdamW optimizer with a learning rate of 1e-5 and a batch size of 32 to update the model parameters for 15 epochs. The evaluation step is set to 100 and all hyperparameters are determined according to the best average model performance at three levels on the validation set. All experiments are performed five times with different random seeds and all reported results are averaged performance.

5.4 Results

Multi-label Classification Comparison The primary experimental results are presented in Table 1, which enables us to draw the following conclusions:

- Firstly, our GOLF model has achieved new state-of-the-art performance across all three levels, as evidenced by both macro-F1 and accuracy metrics. Specifically, on PDTB 2.0, GOLF (base) outperforms the current state-of-the-art LDSGM model (Wu et al., 2022) by 2.03%, 1.25%, and 1.11% in three levels, respectively, in terms of macro-F1. Additionally, it exhibits 1.34%, 0.83%, and 0.65% improvements over the current best results in terms of accuracy. Moreover, in the case of PDTB 3.0, GOLF (base) also outperforms the current state-of-the-art ConnPrompt model (Xiang et al., 2022b) by 1.37% F1 and 1.19% accuracy at the top level.

- Secondly, employing RoBERTa-large embeddings in GOLF leads to a significant improvement in its performance. This observation indicates that our GOLF model can effectively benefit from larger pre-trained language models (PLMs).

- Finally, despite the impressive performance of recent large language models (LLMs) such as ChatGPT (OpenAI, 2022) in few-shot and zero-shot learning for various understanding and reasoning tasks (Bang et al., 2023; Jiang et al., 2023), they still lag behind our GOLF (base) model by approximately 30% in PDTB 2.0. This difference suggests that ChatGPT may struggle to comprehend the abstract sense

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 (%).

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.

of each discourse relation and extract the relevant language features from the text. Therefore, implicit discourse relation recognition remains a challenging and crucial task for the NLP community, which requires further exploration.

Label-wise Classification Comparison Here we present an evaluation of GOLF's performance on PDTB 2.0 using label-wise F1 comparison for top-level and second-level senses. Table 2 showcases the label-wise F1 comparison for the top-level senses, demonstrating that GOLF significantly improves the performance of minority senses such as *Temp* and *Comp*. In Table 3, we compare GOLF with the current state-of-the-art models for the second-level senses. Our results show that GOLF (base) enhances the F1 performance of most second-level senses, with a notable increase in *Expa.List* from 8.98% to 27.78%. Furthermore, by using RoBERTa-large as embeddings, our GOLF (large) model breaks the bottleneck of previous work in two few-shot second-level senses, *Temp.Synchrony* and *Comp.Concession*. To further validate our model's ability of deriving better discourse relation representations, we compare the generated representations of GOLF with those of current state-of-the-art models for both top-level and second-level senses in Appendix B.

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.

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.

Multi-level Consistency Comparison Following (Wu et al., 2022), we evaluate the consistency among multi-level sense predictions via two metrics: 1) Top-Sec: the percentage of correct predictions at both the top-level and second-level senses; 2) Top-Sec-Con: the percentage of correct predictions across all three level senses. Our model's results, as displayed in Table 5, demonstrate more consistent predictions than existing state-of-the-art models in both Top-Sec and Top-Sec-Con, verifying the effectiveness of our model in integrating global and local hierarchical information.

6 Ablation Study

Firstly, we investigate the efficacy of individual modules in our framework. For this purpose, we remove the Multi-Head Interactive Attention (MHIA), the "staircase" in Classifier, the Global Hierarchy-aware Contrastive loss $\mathcal{L}_G$, and the Local Hierarchy-aware Contrastive loss $\mathcal{L}_L$ from GOLF one by one. Note that removing the "staircase" in Classifier means that we keep the cross-entropy loss but remove the dependence between logits from different hierarchical levels. Table 4 indicates that eliminating any of the four modules would hurt the performance across all three levels and reduce the consistency among multi-level label predictions. At the same time, the Local Hierarchy-aware Contrastive loss contributes mostly. Besides, removing both the Global Hierarchy-aware Contrastive loss $\mathcal{L}_G$ and the Local Hierarchy-aware Contrastive loss $\mathcal{L}_L$ significantly hurts the performance. The results show that incorporating label hierarchies from both the global and local perspectives is indeed beneficial. Secondly, we replace the Local Hierarchy-aware Contrastive loss $\mathcal{L}_L$ (Equation (8)) with the hard-label version $\mathcal{L}_{L'}$ (Equation (6)) and find that the performance drops notably. It verifies the usefulness of the scoring function in Equation 7, which considers more subtle semantic structures of local hierarchy. In Appendix C, We also analyze the effects of various hyperparameters consisting of the number layer of MHIA and GCN, the coefficients λ_1 and λ_2, and the temperature τ.

7 Conclusion

In this paper, we present a novel Global and Local Hierarchy-aware Contrastive Framework for implicit discourse relation recognition (IDRR). It can sufficiently incorporate global and local hierarchies to learn better discourse relation representations with the aid of multi-task learning and contrastive learning. Compared with current state-of-the-art approaches, our model empirically reaches better performance at all hierarchical levels of the PDTB dataset and achieves more consistent predictions on multi-level senses.

Limitations

In this section, we illustrate the limitations of our method, which could be summarized into the following two aspects.

Firstly, since the cumbersome data annotation leads to few publicly available datasets of IDRR tasks, we only conduct experiments on English corpora including PDTB 2.0 and PDTB 3.0. In the future, we plan to comprehensively evaluate our model on more datasets and datasets in other languages.

Secondly, considering that instances of PDTB are contained in paragraphs of the Wall Street Journal articles, our approach ignores wider paragraph-level contexts beyond the two discourse arguments. As shown in (Dai and Huang, 2018), positioning discourse arguments in their wider context of a paragraph may further benefit implicit discourse relation recognition. It is worth exploring how to effectively build wider-context-informed discourse relation representations and capture the overall discourse structure from the paragraph level.

Ethics Statement

Since our method relies on pre-trained language models, it may run the danger of inheriting and propagating some of the models' negative biases from the data they have been pre-trained on (Bender et al., 2021). Furthermore, we do not see any other potential risks.

Acknowledgments

W. Wang was supported by HKUST(GZ) Grant G0101000028, GZU-HKUST Joint Research Collaboration Grant GZU22EG04, and Guangzhou Municipal Science and Technology Project (No. 2023A03J0003).

References

Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. 2021. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023a. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *CoRR*, abs/2305.03973.

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

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: contrastive instance learning framework for distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6191–6200. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pages 615–621. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 141–151. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2976–2987.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fengyu Guo, Ruifang He, Jianwu Dang, and Jian Wang. 2020. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7822–7829.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Wu Hong, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2022. Sentence-aware contrastive learning for open-domain passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1062–1074.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Trans. Assoc. Comput. Linguistics*, 3:329–344.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of closed-source large language model. *CoRR*, abs/2305.12870.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Improved universal sentence embeddings with prompt-based contrastive learning and energy-based learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3021–3035, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1152–1158. European Language Resources Association.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3830–3836. ijcai.org.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1224–1233. The Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 10704–10716. Association for Computational Linguistics.

Linh The Nguyen, Ngo Van Linh, Khoat Than, and Thien Huu Nguyen. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence,* *Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4201–4207. Association for Computational Linguistics.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning: Shared Task, CoNLL 2016, Berlin, Germany, August 7-12, 2016*, pages 70–77. ACL.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2339–2352. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics, IWCS 2019, Long Papers, Gothenburg, Sweden, May 23-27 May, 2019*, pages 188–199. Association for Computational Linguistics.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 732–742. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Max Welling and Thomas N Kipf. 2016. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11486–11494.

Changxing Wu, Chaowen Hu, Ruochen Li, Hongyu Lin, and Jinsong Su. 2020. Hierarchical multi-task learning with CRF for implicit discourse relation recognition. *Knowl. Based Syst.*, 195:105637.

Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. 2022a. Encoding and fusing semantic connection and linguistic evidence for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3247–3257, Dublin, Ireland. Association for Computational Linguistics.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022b. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235.

A Data Statistics

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.

B Visualization of Discourse Relation Representations

Here we investigate the quality of discourse relation representations generated by our GOLF model with visualization aids. Figure 4 depicts the 2D t-SNE (Van der Maaten and Hinton, 2008) visualization of discourse relation representations for top-level and second-level senses on the PDTB 2.0 test set. As we can see, compared with current state-of-the-art models BMGF-RoBERTa (Liu et al., 2020) and LDSGM (Wu et al., 2022), our model can generate more centralized discourse relation representations belonging to the same senses (*e.g.*, *Temporal* at the top level, marked in red), and more separated representations belonging to different senses. It verifies our model's capability of deriving better discourse relation representations.

C Effects of Hyperparameters

Here we investigate the effects of various hyperparameters on the development set of PDTB 2.0. These hyperparameters include the number layer L_1 of MHIA (Figure 5), the number layer L_2 of GCN (Figure 6), the coefficient λ_1 of the global hierarchy-aware contrastive loss (Figure 7), the coefficient λ_2 of the local hierarchy-aware contrastive loss (Figure 8), and the temperature τ in contrastive learning (Figure 9). Note that we only change one hyperparameter at a time.

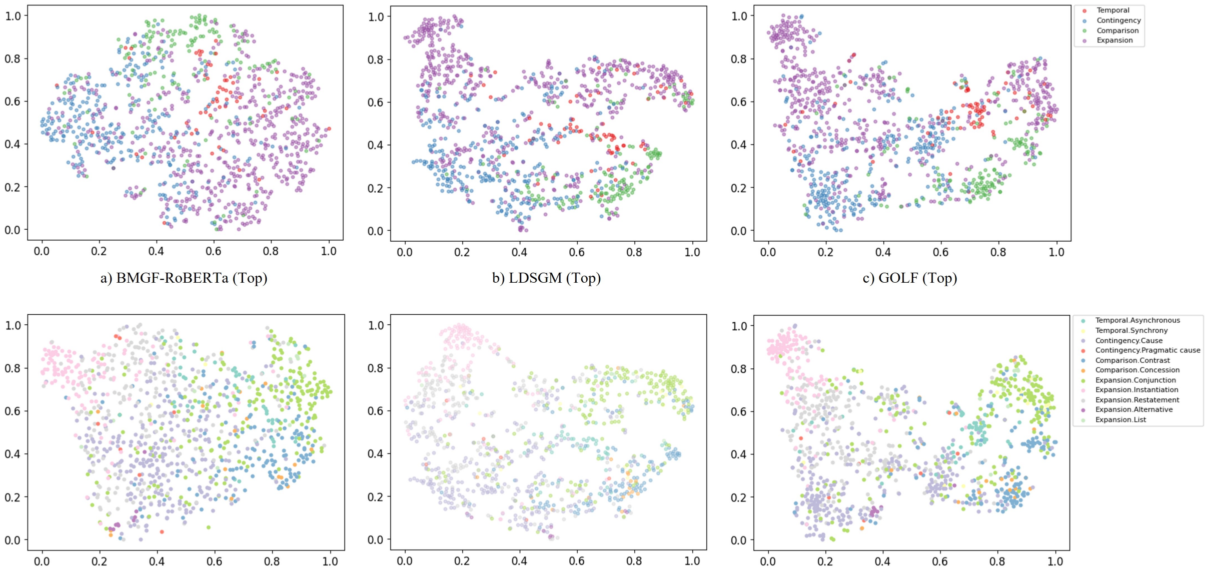
D Label-wise Classification on PDTB 3.0

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**

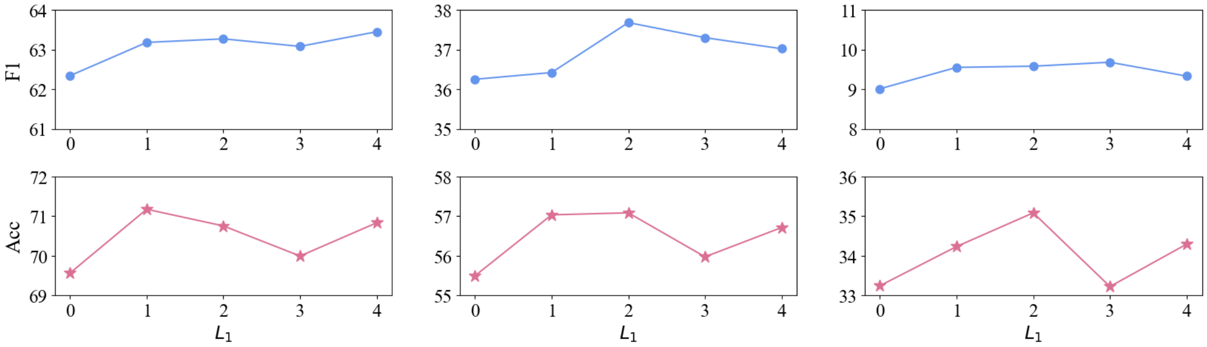
Table 9: Label-wise F1 scores (%) for the second-level senses of PDTB 3.0. The proportion of each sense is listed behind its name.



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

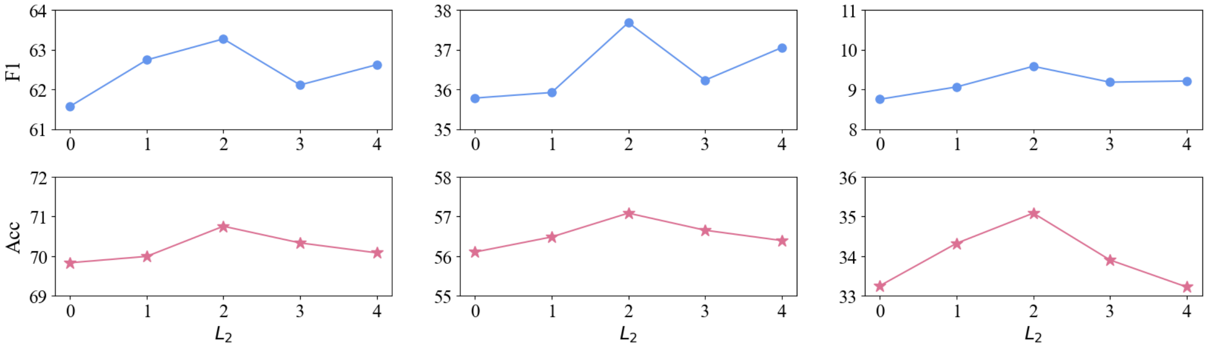
d) BMGF-RoBERTa (Sec) e) LDSGM (Sec) f) GOLF (Sec)

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



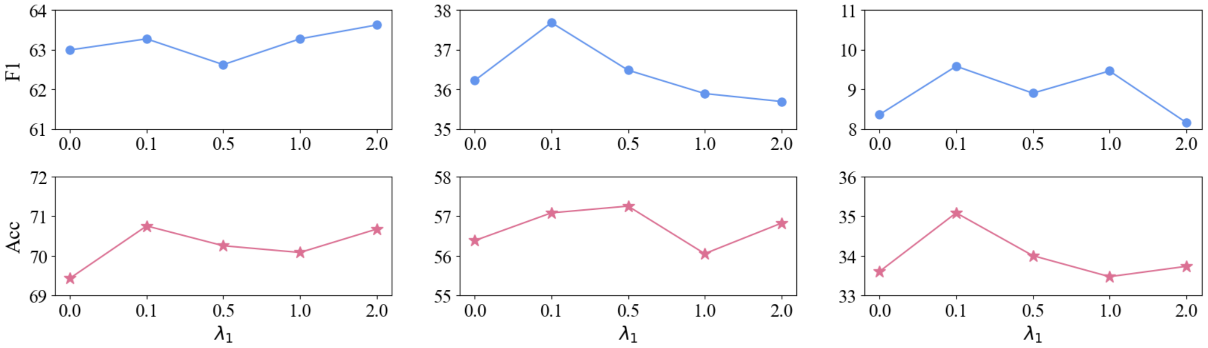
a) Top-level sense classification b) Second-level sense classification c) Implicit connective classification

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



a) Top-level sense classification b) Second-level sense classification c) Implicit connective classification

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

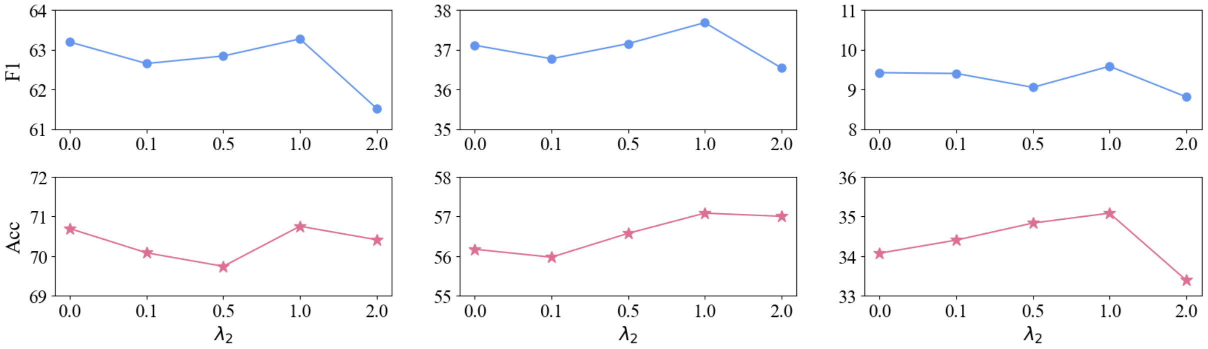


a) Top-level sense classification

b) Second-level sense classification

c) Implicit connective classification

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

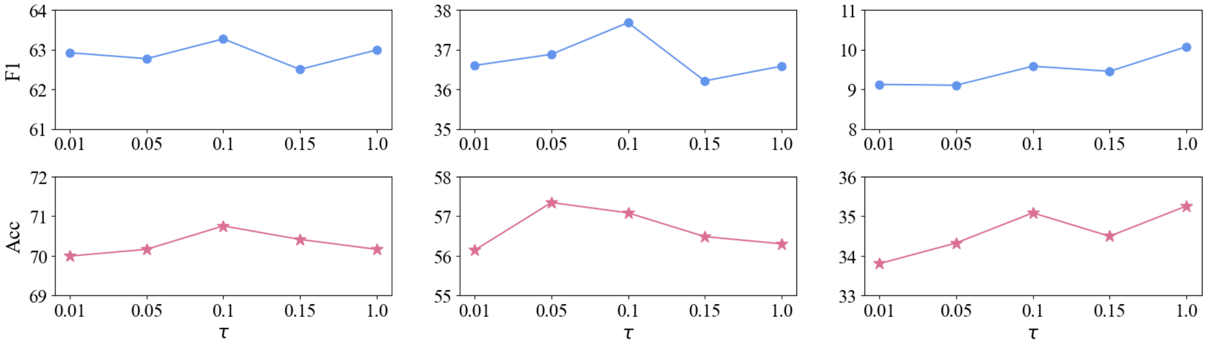


a) Top-level sense classification

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.



a) Top-level sense classification

b) Second-level sense classification

c) Implicit connective classification

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

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The Section named "Limitations"

☑ A2. Did you discuss any potential risks of your work?
The Section named "Ethics Statement"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In the Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In Section 5.3 and 5.4

☑ B1. Did you cite the creators of artifacts you used?
In Section 5.3 and 5.4

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section 5.1 and Appendix A

C ☑ Did you run computational experiments?

In Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 5.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 5.3 and Appendix B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section 5.3 and 5.4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Section 5.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.