Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations

Wanqiu Long[†] and **Bonnie Webber**[†]
[†] University of Edinburgh, Edinburgh, UK
Wanqiu.long@ed.ac.uk, Webber.Bonnie@ed.ac.uk

Abstract

Implicit discourse relation recognition is a challenging task that involves identifying the sense or senses that hold between two adjacent spans of text, in the absense of an explicit connective between them. In both PDTB-2 (Prasad et al., 2008) and PDTB-3 (Webber et al., 2019), discourse relational senses are organized into a three-level hierarchy ranging from four broad top-level senses, to more specific senses below them. Most previous work on implicitf discourse relation recognition have used the sense hierarchy simply to indicate what sense labels were available. Here we do more — incorporating the sense hierarchy into the recognition process itself and using it to select the negative examples used in contrastive learning. With no additional effort, the approach achieves state-of-the-art performance on the task. Our code is released in https://github.com/wanqiulong 0923/Contrastive_IDRR.

1 Introduction

Discourse relations are an important aspect of textual coherence. In some cases, a speaker or writer signals the sense or senses that hold between clauses and/or sentences in a text using an explicit connective. Recognizing the sense or senses that hold can be more difficult, in the absense of an explicit connective.

Automatically identifying the sense or senses that hold between sentences and/or clauses can be useful for downstream NLP tasks such as text summarization (Cohan et al., 2018), machine translation (Meyer et al., 2015) and event relation extraction (Tang et al., 2021). Recent studies on implicit discourse relation recognition have shown great success. Especially, pre-trained neural language models (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019) have been used and dramatically improved the performances of models (Shi and Demberg, 2019b; Liu et al., 2020; Kishimoto et al., 2020). The senses available for labelling
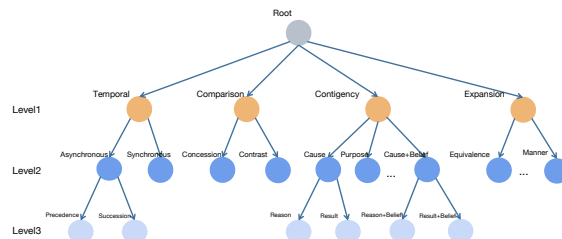

Figure 1: The PDTB-3 Sense Hierarchy

discourse relations in the PDTB-2 (and later in the PDTB-3) are arranged in a three-level hierarchy, with the most general senses at the top and more specific senses further down. In the PDTB-3, annotators could only choose senses at terminal nodes in the hierarchy – level-2 senses for symmetric relations such as EXPANSION.EQUIVALENCE and TEMPORAL.SYNCHRONOUS, and level-3 senses for asymmetric relations, with the direction of the relation encoded in its sense label such as SUBSTITUTION.ARG1-AS-SUBST (where the text labelled ARG1 substitutes for the denied text labelled ARG2) and SUBSTITUTION.ARG2-AS-SUBST (where the text labelled ARG2 substitutes for the denied text labelled ARG1). Early work on recognizing the implicit relations only used the hierarchy to choose a target for recognition (e.g., the senses at level-1 (classes) or those at level-2 (types). Recently, Wu et al. (2022) have tried to leverage the dependence between the level-1 and level-2 labels (cf. Section 2). The current work goes further, using the whole three-level sense hierarchy to select the negative examples for contrastive learning.

Contrastive learning, which aims to minimize the distance between similar instances (defined as positive examples) and widen the difference with dissimilar instances (negative examples), has been considered as effective in constructing meaningful representations (Kim et al., 2021; Zhang et al., 2021; Yan et al., 2021). Previous work on contrastive learning indicates that it is critical to select good negative samples (Alzantot et al., 2018;

10704

Wu et al., 2020b; Wang et al., 2021). The insight underlying the current work is that the hierarchy of sense labels can enable the selection of good negative examples for contrastive learning. To see this, consider Examples 1-3 below from the PDTB-3. On the surface, they look somewhat similar, but in Examples 1 and 2, the annotators took the second sentence (Arg2) as providing more detail about the first sentence (Arg1) — the sense called EXPANSION.LEVEL-OF-DETAIL.ARG2-AS-DETAIL, while in Example 3, they took the second sentence as expressing a substitute for "American culture" in terms of what is relevant – the sense called EXPANSION.SUBSTITUTION.ARG2-AS-SUBST.

(1) *"Valley National ""isn't out of the woods yet "*. **The key will be whether Arizona real estate turns around or at least stabilizes.**.

(2) *The House appears reluctant to join the senators*. **A key is whether House Republicans are willing to acquiesce to their Senate colleagues' decision to drop many pet provisions.**.

(3) *Japanese culture vs. American culture is irrelevant*. **The key is how a manager from one culture can motivate employees from another.**.

In this work, we use a multi-task learning framework, which consists of classification tasks and a contrastive learning task. Unlike most previous work using one benchmark dataset (usually PDTB-2 or PDTB-3), we evaluate our systems on both PDTB-2 and PDTB-3. Besides, Wang et al. (2021) have shown that data augmentation can make representations be more robust, thereby enriching the data used in training. We thus follow Ye et al. (2021) and Khosla et al. (2020) in identifying a relevant form of data augmentation for our contrastive learning approach to implicit relation recognition.

The main contributions of our work are as follows:

- We leveraged the sense hierarchy to get contrastive learning representation, learning an embedding space in which examples from same types at level-2 or level-3 stay close to each other while sister types are far apart.
- We explored and compared different methods of defining the negatives based on the sense

hierarchies in PDTB-2 and PDTB-3, finding the approach which leads to the greatest improvements.
- Our proposed data augmentation method to generate examples is helpful to improve the overall performance of our model.
- We demonstrate that implicit relation recognition can benefit from a deeper understanding of the sense labels and their organization.

2 Related Work

Implicit discourse relation recognition For this task, Dai and Huang (2018) considered paragraph-level context and inter-paragraph dependency. Recently, Shi and Demberg (2019b) showed that using the bidirectional encoder representation from BERT (Devlin et al., 2019) is more accurately to recognize Temporal.Synchrony, Comparison.Contrast, Expansion.Conjunction and Expansion.Alternative. Liu et al. (2020) showed that different levels of representation learning are all important to implicit relation recognition, and they combined three modules to better integrate context information, the interaction between two arguments and to understand the text in depth. However, only two existing works leveraged the hierarchy in implicit relation recognition. Both Wu et al. (2020a) and Wu et al. (2022) first attempted to assign a Level-1 sense that holds between arguments, and then only considered as possible Level-2 senses, those that are daughters of the Level-1 sense.

Contrastive learning Recently, there has been a growing interest in applying contrastive learning in both the pre-training and fine-tuning objectives of pre-trained language models. Gao et al. (2021) used a contrastive objective to fine-tune pre-trained language models to obtain sentence embeddings, and greatly improves state-of-the-art sentence embeddings on semantic textual similarity tasks. Suresh and Ong (2021) proposed label-aware contrastive loss in the presence of larger number and/or more confusable classes, and helps models to produce more differentiated output distributions. Besides, many works have demonstrated that selecting good negative examples are very important for using contrastive learning (Schroff et al., 2015; Robinson et al., 2021; Cao et al., 2022). In our work, we integrate contrastive learning loss with supervised losses and we use the structure of the sense hierarchy to guide the selection of negative

examples.

3 Learning Loss

3.1 Supervised Learning Loss

The standard approach today for classification task is to use a standard cross-entropy loss:

$$\mathcal{L}_{sup} = \frac{1}{N} \sum_{i=1}^{N} -log \frac{e^{W_{y_i}^T s_i}}{\sum_j e^{W_j^T s_i}} \qquad (1)$$

Where N denotes the number of training examples, y_i is the ground-truth class of the i-th class and W_j is the weight vector of the j-th class.

3.2 Contrastive Learning Loss

In contrastive learning, each example can be treated as an anchor to get its positive and negative examples. Contrastive learning can pull the anchor and its positive example together in the embedding space, while the anchor and negative samples are pushed apart. The contrastive learning loss was used by Chen et al. (2020); Suresh and Ong (2021) before. A set of N randomly sampled label pairs is defined as x_k, y_k, where x and y represent samples and labels, respectively, $k = 1, ..., N$. Let i be the index of anchor sample and j is the index of a positive sample. where $i\epsilon\{1, ..., N\}, i \neq j$. Contrastive loss is defined as:

$$\mathcal{L}_{scl} = -\sum_{i=1}^{N} \frac{e^{sim(h_j,h_i)}\tau}{\sum_{i \neq k} e^{sim(h_k,h_i)}\tau} \qquad (2)$$

Here, h denotes the feature vector in the embedding space, and τ is the temperature parameter. Intuitively, the numerator computes the inner dot product between the anchor points i and its positive sample j. The denominator computes the inner dot product between all i and the inner dot product between all negative samples. where a total of $N - 1$ samples are computed.

Supervised contrastive learning (Gunel et al., 2021) extends the equation.2 to the supervised scenario. In particular, given the presence of labels, the positive examples are all examples with the same label. The loss is defined as:

$$\mathcal{L}_{scl} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} 1_{i \neq j} 1_{y_i = y_j}$$
$$log \frac{e^{sim(h_j,h_i)\tau}}{\sum_{k=1}^{N} 1_{i \neq k} e^{sim(h_k,h_i)/\tau}} \qquad (3)$$

N_{y_j} indicates the number of examples in a batch that have the same label as i, τ is the temperature parameter and h denotes the feature vector that is from the l2 normalized final encoder hidden layer before the softmax projection.

4 Our Approach

Figure 2 shows the overall architecture of our method. As figure 2 illustrates, we firstly use a simple multi-task model based on RoBERTa-base (Liu et al., 2019), and then we develop a contrastive learning algorithm where the sense hierachy is used to select positive and negative examples. Detailed descriptions of our framework and our data augmentation method are given below.

4.1 Sentence Encoder

Every annotated discourse relation consists of two sentences or clauses (its *arguments*) and one or more relational senses that the arguments bear to each other. We concatenate the two arguments of each example and input them into RoBERTa. Following standard practices, we add two special tokens to mark the beginning ([CLS]) and the end ([SEP]) of sentences. We use the representation of [CLS] in the last layer as the representation of the whole sentences.

4.2 Data Augmentation

To increase the number of training examples, we take advantage of meta-data recorded with each Implicit Discourse Relation in the PDTB (cf. (Webber et al., 2019), Section 8]). For each sense taken to hold between the arguments of that relation, annotators have recorded in the meta-data, an explicit connective that could have signalled that sense. In the past, this meta-data was used in implicit relation recognition by both Patterson and Kehler (2013) and Rutherford and Xue (2015). We have used it in a different way, shown in Figure 3, to create an additional training example for each connective that appears in the meta-data. In the added training example, this added connective becomes part of the second argument of the relation (i.e., appearing after the [SEP] character)

Since there is at least **one** explicit connective recorded in the meta-data for each implicit discourse relation and at most **two** [1], for a training batch of N tokens, there will be at least another

---

[1] This is because the PDTB only allows for one or two senses per relation.
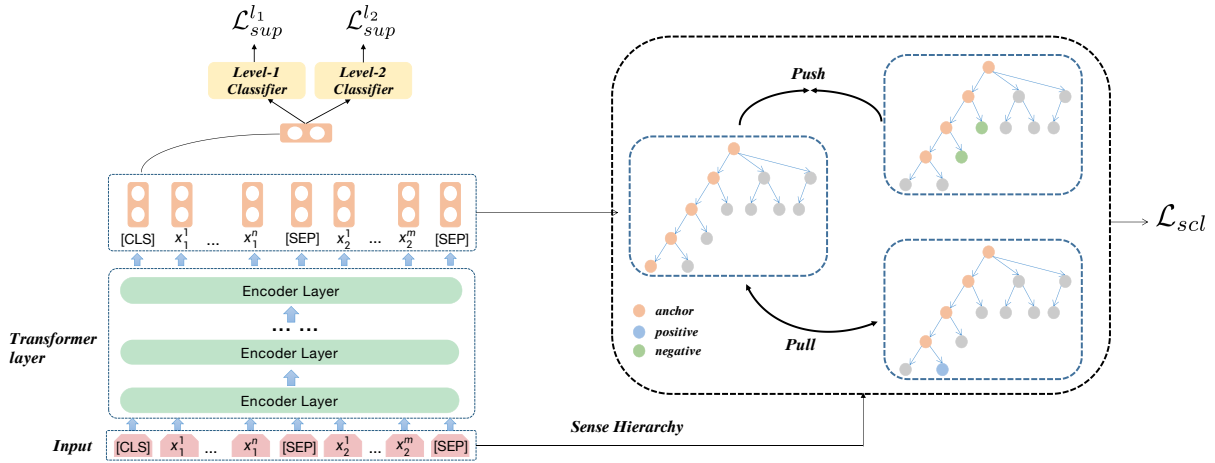
Figure 2: **The overall architecture of our model.** When given an anchor, we search the positive and negative examples in a training batch based on the sense hierarchy of the PDTB. We narrow the distances among examples from the same types at level-2 or level-3 and enlarge the distances among examples from different types at level-2 and level-3.

N tokens introduced by this *data augmentation* method, increasing the training batch to at least 2N tokens.

Figure 3: An example with inserted connective: the connective word is "In contrast".

4.3 Positive Pair and Negative Pair Generation

We use the structure of the sense hierarchy to identify the positive and negative examples needed for contrastive learning. The only senses used in annotating discourse relations are ones at terminal nodes of the sense hierarchy. This is Level 2 for symmetric senses and Level 3 for asymmetric senses (i.e., where the **inverse** of the sense that holds between Arg1 and Arg2 is what holds between Arg2 and Arg1. For example, CONTRAST and SIMILARITY are both symmetric senses, while MANNER and CONDITION are asymmetric, given that there is a difference between Arg2 being the manner of doing Arg1 or Arg1 being the manner of doing Arg2). In our work, when the lowest level of the senses is level-3, we directly used the level-3 labels instead of their parent at level-2. For example, under the level-2 label Temporal.asynchronous, there are two labels which are precedence and succession at level-3. For this case, we replaced the level-2 label Temporal.asynchronous with the two labels precedence and succession at level-3.

Although supervised contrastive learning in Eq. 3 can be valid for different classes of positive ex-

ample pairs, its negative examples come from any examples inside a batch except itself. We defined l_1, l_2, l_3 as the first, second, and third level in the hierarchical structure respectively, and $l \epsilon l_i$ refers to the labels from level i.

Instance $e \sim$ Same sub-level $\quad e_{pos}$
Given the representation of a sentence e_i and its first, second and third level of label l_1^i, l_2^i, l_3^i, we searched the set of examples with the same second level labels or the same third level labels (if the lowest level is level-3) as e_{pos} in each training batch:

$$e_{pos}^i = \{e \in e_{pos}^i : l_2^e == l_2^i \quad or \quad l_3^e == l_3^i\} \tag{4}$$

E.g. If the label of the anchor is Temporal.asynchronous.precedence, its positive examples would be the examples with the same label.

Instance $e \sim$ Batch instance $\quad e_{neg}$
Here, we would like to help the model discriminate the sister types at level-2 and level-3 (if the lowest level is level-3). We searched the set of examples with different level-2 labels or level-3 labels as e_{neg} in each training batch.

E.g. If the label of the anchor is Temporal.asynchronous.precedence, its negative examples would be its sister types at level-2 and level-3, namely Temporal.asynchronous.succession and Temporal.synchronous.

$$\begin{aligned} e_{neg}^i = \{e \in e_{neg}^i : l_1^e == l_1^i \quad \& \\ (l_2^e \neq l_2^i \quad \& \quad l_3^e \neq l_3^i)\} \end{aligned} \tag{5}$$

4.4 Loss Algorithms

As described above, given the query e_i with its positive pairs and negative pairs and based on the general contrastive learning loss (see Equation 2), the contrastive learning loss for our task and approach is:

$$\mathcal{L}_{scl} = \sum_{i=1}^{N} -\frac{1}{|e_{pos}^i| - 1} \sum_{j=1}^{2N} 1_{i \neq j} 1_{j \in e_{pos}^i}$$

$$log \frac{w_j e^{sim(h_j, h_i)\tau}}{\sum_{k=1}^{2N} 1_{i \neq k} 1_{k \in e_{neg}^i + e_{pos}^i} w_k e^{sim(h_k, h_i)/\tau}} \quad (6)$$

where w_j and w_j are weight factors for different positive pairs and negative pairs respectively, $sim(h_i, h_j)$ is cosine similarity and τ is a temperature hyperparameter.

Our overall training goal is:

$$\mathcal{L} = \mathcal{L}_{sup}^{l1} + \mathcal{L}_{sup}^{l2} + \beta \mathcal{L}_{scl} \quad (7)$$

As our classifications are done in the first level and second level for the same inputs, we used a standard cross-entropy loss to get supervised loss $\mathcal{L}_{sup}^{L1}$ and $\mathcal{L}_{sup}^{L2}$. And β is the weighting factor for the contrastive loss.

5 Experiment Setting

5.1 Datasets

Besides providing a sense hierarchy, the Penn Discourse TreeBank (PDTB) also frequently serves as a dataset for evaluating the recognition of discourse relations. The earlier corpus, PDTB-2 (Prasad et al., 2008) included 40,600 annotated relations, while the later version, PDTB-3 (Webber et al., 2019) includes an additional 13K annotations, primarily intra-sentential, as well as correcting some inconsistencies in the PDTB-2. The sense hierarchy used in the PDTB-3 differs somewhat from that used in the PDTB-2, with additions motivated by the needs of annotating intra-sentential relations and changes motivated by difficulties that annotators had in consistently using some of the senses in the PDTB-2 hierarchy.

Because of the differences in these two hierarchies, we use the PDTB-2 hierarchy for PDTB-2 data and the PDTB-3 hierarchy for PDTB-3 data respectively. We follow earlier work (Ji and Eisenstein, 2015; Bai and Zhao, 2018; Liu et al., 2020; Xiang et al., 2022) using Sections 2-20 of the corpus for Training, Sections 0-1 for Validation, and Sections 21-22 for testing. We followed the predecessors to split section 2-20, 0-1, and 21-22 as training, validation, and test sets for both PDTB-2 and PDTB-3. With regard to those instances with multiple annotated labels, we also follow previous work (Qin et al., 2016). They are treated as separate examples during training. At test time, a prediction matching one of the gold types is taken as the correct answer. Implicit relation recognition is usually treated as a classification task. While 4-way (Level-1) classification was carried out on both PDTB-2 and PDTB-3, more detailed 11-way (Level 2) classification was done only on the PDTB-2 and 14-way (Level 2) classification, only on the PDTB-3.

5.2 Baselines

To exhibit the effectiveness of our proposed method, we compare our method with strong baselines. As previous work usually used one dataset (PDTB-2 or PDTB-3) for evaluation, we use different baselines for PDTB-2 and PDTB-3. Since PDTB-3 was not released until 2019, the baselines for PDTB-3 from 2016 and 2017 are from (Xiang et al., 2022). They reproduced those models which were originally used on PDTB-2 on PDTB-3.

Baselines for PDTB-2:

- (Dai and Huang, 2019): a neural model leveraging external event knowledge and coreference relations.
- (Shi and Demberg, 2019a): a neural model that leverages the inserted connectives to learn better argument representations.
- (Nguyen et al., 2019): a neural model which predicts the labels and connectives. simultaneously.
- (Guo et al., 2020): a knowledge-enhanced Neural Network framework.
- (Kishimoto et al., 2020): a model applying three additional training tasks.
- (Liu et al., 2020): a RoBERTa-based model which consists of three different modules.
- (Jiang et al., 2021): a method that recognizes the relation label and generates the target sentence simultaneously.
- (Dou et al., 2021): a method using conditional VAE to estimate the risk of erroneous sampling.
- (Wu et al., 2022): a label dependence-aware sequence generation model.
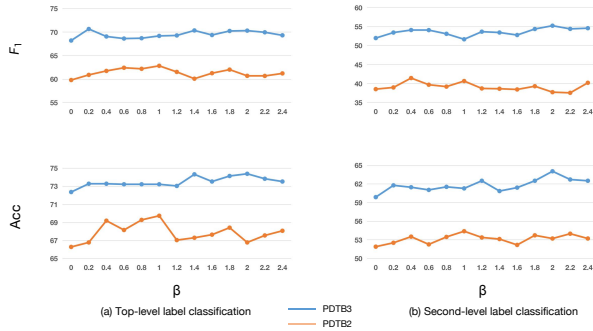
Baselines for PDTB-3:

Figure 4: Effects of β on the validation set.

- (Liu and Li, 2016): a model that combines two arguments' representation for stacked interactive attention.
- (Chen et al., 2016a): a mixed generative-discriminative framework.
- (Lan et al., 2017): a multi-task attention neural network.
- (Ruan et al., 2020): a propagative attention learning model.
- (Xiang et al., 2022): a model that uses a Dual Attention Network (DAN).

5.3 Parameters Setting

In our experiments, we use the pre-trained RoBERTa-base (Liu et al., 2019) as our Encoder. We adopt Adam (Kingma and Ba, 2015) with the learning rate of $3e-5$ and the batch size of 256 to update the model. The maximum training epoch is set to 25 and the wait patience for early stopping is set to 10 for all models. We clip the gradient L2-norm with a threshold 2.0. For contrast learning, the weight of positive examples is set to 1.6 and the weight of negative examples is set to 1. All experiments are performed with 1× 80GB NVIDIA A100 GPU.

5.4 Evaluation Metrics

We used Accuracy and Macro-F1 score as evaluation metrics, because PDTB datasets are imbalanced and Macro-F1 score has been said to be an more appropriate assessment measure for imbalanced datasets (Akosa, 2017; Bekkar et al., 2013).

5.5 Effects of the Coefficient β

As shown in Equation 7, the coefficient β is an important hyperparameter that controls the relative importance of supervised loss and contrastive loss. Thus, we vary β from 0 to 2.4 with an increment of 0.2 each step, and inspect the performance of our model using different β on the validation set.

From Figure 4, we can find that, compared with the model without contrastive learning ($\beta = 0$), the performance of our model at any level is always improved via contrastive learning. For PDTB-2, when β exceeds 1.0, the performance of our model tends to be stable and declines finally. Thus, we directly set $\beta = 1.0$ for all PDTB-2 related experiments thereafter. For PDTB-3, the Acc and F1 of the validation set reach the highest point at $\beta = 2.0$. Therefore we choose $\beta = 2.0$ for all related experiments.

We have considered three ways of investigating why there is such a difference in the optimal weighting coefficient. First, compared with PDTB-2, the PDTB-3 contains about 6000 more implicit tokens annotated for discourse relations. Secondly, although the sense hierarchies of both the PDTB-2 and the PDTB-3 have three levels and have the same senses at level-1, but many changes at level-2 and level-3 due to difficulties found in annotating certain senses. Moreover, the intra-sentential implicit relations might be another reason. In PDTB-3, many more discourse relations are annotated within sentences. Liang et al. (2020) report quite striking difference in the distribution of sense relations inter-sententially vs. intra-sententially between PDTB-2 and PDTB-3. Therefore, these major differences in the PDTB-3 and the PDTB-2 might cause the fluctuation of the coefficient value.

6 Results and Analysis

The results on PDTB-2 and PDTB-3 for Level-1 and Level-2 are presented in Table 1 and Table 2 respectively, where the best results are highlighted in bold. Classification performance on PDTB-2 in terms of Macro-F1 for the four general sense types at Level-1 and 11 sense types at Level-2 is shown in Table 3 and Table 4.

These results demonstrate better performance than previous systems for both Level-1 and Level-2 classification on both PDTB-2 and PDTB-3. In particular, the results clearly demonstrate benefits to be gained from contrastive learning. But there is more to be said: In Section 6.1, we discuss different ways of defining negative examples with respect to the sense hierarchy, and in Section 6.2, we discuss the relative value of the particular form of data augmentation we have used (cf. Section 4.2) as compared with our method of contrastive learning.

	PDTB-2			
Model	Top Level		Second Level	
	Acc	Macro-F1	Acc	Macro-F1
Dai and Huang (2019)	59.66	52.89	48.23	33.41
Shi and Demberg (2019a)	61.42	46.40	47.83	-
Nguyen et al. (2019)	-	53.00	49.95	-
Guo et al. (2020)	57.25	47.90	-	-
Kishimoto et al. (2020)	65.26	58.48	52.34	-
Liu et al. (2020)	69.06	63.39	58.13	-
Jiang et al. (2021)	-	57.18	-	37.76
Dou et al. (2021)	70.17	65.06	-	-
Wu et al. (2022)	71.18	63.73	60.33	40.49
Ours	**72.18**	**69.60**	**61.69**	**49.66**

Table 1: Experimental results on PDTB-2.

	PDTB-3			
Model	Top Level		Second Level	
	Acc	Macro-F1	Acc	Macro-F1
Liu and Li (2016)	57.67	46.13	-	-
Chen et al. (2016b)	57.33	45.11	-	-
Lan et al. (2017)	57.06	47.29	-	-
Ruan et al. (2020)	58.01	49.45	-	-
Xiang et al. (2022) (BiLSTM)	60.45	53.14	-	-
Xiang et al. (2022) (BERT)	64.04	56.63	-	-
Ours	**75.31**	**70.05**	**64.68**	**57.62**

Table 2: Experimental results on PDTB-3.

Model	Comp.	Cont	Exp.	Temp.
Nguyen et al. (2019)	48.44	56.84	73.66	38.60
Guo et al. (2020)	43.92	57.67	73.45	36.33
Liu et al. (2020)	59.44	60.98	77.66	50.26
Jiang et al. (2021)	55.40	57.04	74.76	41.54
Dou et al. (2021)	55.72	63.39	**80.34**	44.01
Ours	**65.84**	**63.55**	79.17	**69.86**

Table 3: The results for relation types at level-1 on PDTB-2 in terms of F1 (%) (top-level multi-class classification).

Second-level Label	Liu et al. (2020)	Wu et al. (2022)	ours
Temp.Asynchronous	56.18	56.47	**59.79**
Temp.Synchrony	0.00	0.00	**78.26**
Cont.Cause	59.60	64.36	**65.58**
Cont.Pragmatic cause	0.0	0.0	0.00
Comp.Contrast	59.75	**63.52**	62.63
Comp.Concession	0.00	0.00	0.00
Exp.Conjunction	**60.17**	57.91	58.35
Exp.Instantiation	67.96	72.60	**73.04**
Exp.Restatement	53.83	58.06	**60.00**
Exp.Alternative	60.00	**63.46**	53.85
Exp.List	0.0	8.98	**34.78**

Table 4: The results for relation types at level-2 on PDTB-2 in terms of F1 (%) (second-level multi-class classification).

6.1 Comparisons with Other Negatives Selecting Methods

There is not only one way to select negative examples for contrastive learning based on PDTB hierarchical structures. In addition to the method we adopt, we have explored another 4 different methods of defining positive and negative examples by using the sense hierarchies, which can be shown in Figure 5. One can choose the level against which to select negative examples: method 2 below uses examples with different labels at level-2, while methods 1, 3 and 4 use examples with different labels at level-1. With regard to the use of weight for method 3 and method 4, we aim to give more weight to more similar (potentially) positive examples based on the hierarchy. Specifically, we give more weight to the examples from the same level-2/level-3 type than their sister types at level-2/level-3 when all of the examples from the same level-1 are positive examples. Besides, method 4 leverages level-3 labels, while method 1 to 3 only consider level-1 and level-2 labels. In our experiments for other negatives defining methods, we use the same hyperparameters as the experimental setup of our methods. For method 3 and method 4, the weight of positive examples is set to 1.6 and 1.3 and the weight of negative examples still is 1.

It can be seen from table 5 and table 6 that our method is better than the above methods in both

datasets for both level-1 and level-2 classification tasks. Compared with method 2, we utilize level-3 labels, which indicated the level-3 label information is helpful for the approach. The greatest difference between our method and other three methods is that our negative examples are only those sister types at level-2 or level-3, not including the examples from different level-1. On the contrary, the negative examples in those three methods are examples from other level-1 types. We suppose that this might make a too strong assumption that examples from different level-1 are very dissimilar. In PDTB datasets, some examples have been annotated with multiple labels. We found that among all examples with multiple annotated labels, there are 99.26% examples whose multiple labels are under different level-1. Moreover, some level-1 types of relation might be overlapped even if the annotators just annotate one label. For example, some examples annotated as Temporal.asynchronous might have the sense of Contingency.cause as well. And Moens and Steedman (1988) have pointed out that *when*-clauses do not simply predicate a temporal relation, but a causal one as well, which can be called contingency. This shows up in the PDTB in terms of the variation in how particular tokens of *when* clauses have been annotated. But it also means that in choosing Negative examples, relations labelled TEMPORAL.SYNCHRONOUS or TEMPORAL.ASYNCHRONOUS may closely resemble those labelled CONTINGENCY.CAUSE and therefore not

Model	PDTB-2				PDTB-3			
	Top Level		Second Level		Top Level		Second Level	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Method 1	68.91	65.04	58.61	46.27	73.25	68.00	61.17	55.58
Method 2	69.39	63.95	58.33	44.80	73.53	68.36	61.93	54.85
Method 3	69.39	66.53	58.61	39.20	72.49	67.49	60.77	54.33
Method 4	69.10	65.30	57.07	47.46	71.26	66.47	59.53	47.24
Ours	**72.18**	**69.60**	**61.69**	**49.66**	**75.31**	**70.05**	**64.48**	**57.62**

Table 5: Comparisons with other negatives defining methods.

Positve: examples with same label at level-1.
Negative: examples with different labels at level-1.

(a) method 1

Positve: examples with same label at level-2.
Negative: examples with different labels at level-2.

(b) method 2

Positve: Examples with same label at level-1.
More weight are given to the examples
with same label at level-2.
Negative: Examples with different labels at level-1.

(c) method 3

Positve: Examples with same label at level-1,
more weight are given to the examples
with same label at level-2 or level-3.
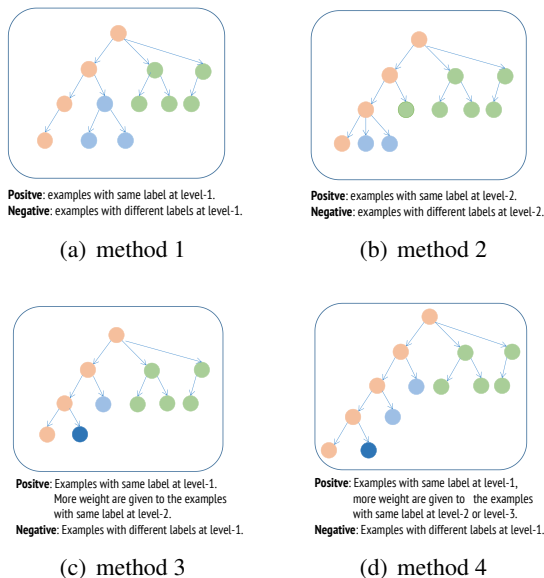Negative: Examples with different labels at level-1.

(d) method 4

Figure 5: The other four negative examples selected methods. orange ball represent anchor, green ball represent negative examples, and blue ball represent positive examples. Darker blue ball means more weight is given to more similar (potentially) positive examples.

be effective as negative examples. Specifically, for the following example:

(4) **when** [they built the 39th Street bridge]$_1$, [they solved most of their traffic problems]$_2$.

If the connective "when" is replaced with "because", the sentence still sounds not strange. Therefore, regarding all examples from different level-1 as negative examples might have some negative impacts on learning the representations.

6.2 Ablation Study

We wanted to know how useful our data augmentation method and our contrastive learning method are, so we have undertaken ablation studies for this.
Effects of contrastive learning algorithm From Table 7, it can be seen that multi-task learning method where level-1 and level-2 labels are predicted simultaneously by using the same [CLS]

Model	Comp.	Cont	Exp.	Temp.
Method 1	63.26	60.42	76.78	59.74
Method 2	60.78	60.82	77.89	56.30
Method 3	59.85	**65.18**	76.43	64.67
Method 4	57.25	61.73	77.30	64.90
Ours	**65.84**	63.55	**79.17**	**69.86**

Table 6: The results of relation types at level-1 on PDTB-2 in terms of F1 (%) (top-level multi-class classification).

representation perform better than separately predicting level-1 and level-2 labels, which verifies the dependency between different levels. Compared with the multi-task learning method, our model with a contrastive loss has better performance in PDTB-2 and PDTB-3, which means that our contrasting learning method is indeed helpful.
Effects of data augmentation Table 8 compares the results with and without data augmentation for both PDTB-2 and PDTB-3. From the comparisons, it is clear that the data augmentation method is helpful to generate useful examples. Khosla et al. (2020) showed that having a large number of hard positives/negatives in a batch leads to better performance. Since we have many classes at the second level, 11 types for PDTB-2 and 14 types for PDTB-3. In a batch with the size of 256, it is difficult to guarantee that there are enough positive examples for each class to take full advantage of contrast learning. Therefore, without data augmentation, the performance of our method degrades considerably.

7 Limitations and Future work

With regard to PDTB-2 and PDTB-3 annotation, there are two cases: (1) Annotators can assign multiple labels to an example when they believe more than one relation holds simultaneously; (2) Annotators can be told (in the Annotation Manual) to give precedence to one label if they take more than one to hold. For example, they are told in

Datasets	Model	Top Level		Second Level	
		Acc	Macro-F1	Acc	Macro-F1
PDTB-2	RoBERTa	68.14	64.87	58.33	48.37
	RoBERTa-MTL	69.87	65.39	58.22	45.21
	Ours	**72.18**	**69.60**	**61.69**	**49.66**
PDTB-3	RoBERTa	72.02	67.44	60.56	57.12
	RoBERTa-MTL	72.63	68.23	60.56	57.16
	Ours	**75.31**	**70.05**	**64.68**	**57.62**

Table 7: Ablation study on PDTB-2 and PDTB-3.

Model		Top Level		Second Level	
		Acc	Macro-F1	Acc	Macro-F1
PDTB-2	Ours	**72.18**	**69.60**	**61.69**	**49.66**
	-augmentation	71.70	67.85	59.19	45.54
PDTB-3	Ours	**75.31**	**70.05**	**64.68**	**57.62**
	-augmentation	73.32	69.02	63.24	51.80

Table 8: Effects of data augmentation.

the Manual (Webber et al., 2019) that examples that satisfy the conditions for both Contrast and Concession, should be labelled as concession. We over-simplified the presence of multiple labels by following Qin et al. (2017) in treating each label as a separate example and did not consider the second case. Thus, our approach might be inadequate for dealing with the actual distribution of the data and can be extended or modified. It is worth exploring how to extend our approach to allow for examples with multiple sense labels and cases where one label takes precedence over another. We believe that this will be an important property of the work.

Another limitation is that we only use English datasets. There are PDTB-style datasets in other languages including a Chinese TED dicourse bank corpus (Long et al., 2020), a Turkish discourse Tree bank corpus (Zeyrek and Kurfalı, 2017) and an Italian Discourse Treebank (Pareti and Prodanof, 2010). Moreover, Zeyrek et al. (2019) proposed a TED Multilingual Discourse Bank (TED-MDB) corpus, which has 6 languages. These datasets allow us to assess the approach in languages other than English. Besides, there are datasets similar to PDTB-Style like Prague Dependency Treebank (Mírovský et al., 2014). The different datasets use essentially similar sense hierarchy, but two things need to be investigated (i) whether there are comparable differences between tokens that realise "sister" relations, or (ii) whether tokens often have multiple sense labels, which would change what could be used as negative examples if leveraging our approach on them.

In the future, we can also assess whether contrastive learning could help in separating out EntRel relations and AltLex relations from implicit relations or whether other methods would perform better.

8 Conclusions

In this paper, we leverage the sense hierarchy to select the negative examples needed for contrastive learning for the task of implicit discourse relation recognition. Our method has better overall performance than achieved by previous systems, and compared with previous work, our method is better at learning minority labels. Moreover, we compared different methods of selecting the negative examples based on the hierarchical structures, which shows some potential negative impacts might be produced when negative examples include those from other level-1 types. Moreover, we conduct ablation studies to investigate the effects of our data augmentation method and our contrastive learning method. Besides, the limitations and the future work are discussed.

Acknowledgments

References

Josephine Sarpong Akosa. 2017. Predictive accuracy : A misleading performance measure for highly imbalanced data.

Moustafa Farid Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *EMNLP*.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, Santa

Fe, New Mexico, USA. Association for Computational Linguistics.

Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. 2013. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3:27–38.

Rui Cao, Yihao Wang, Yu Guo Liang, Ling Gao, Jie Zheng, Jie Ren, and Zheng Wang. 2022. Exploring the impact of negative samples of contrastive learning: A case study of sentence embedding. In *FINDINGS*.

Jifan Chen, Qi Zhang, Pengfei Liu, and Xuanjing Huang. 2016a. Discourse relations detection via a mixed generative-discriminative framework. In *AAAI*.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016b. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1735, Berlin, Germany. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 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, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. 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, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. 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, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zujun Dou, Yu Hong, Yu Sun, and Guodong Zhou. 2021. CVAE-based re-anchoring for implicit discourse relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1275–1283, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

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 *AAAI*.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online. Association for Computational Linguistics.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. In *LREC*.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark. Association for Computational Linguistics.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.

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-20*, pages 3830–3836. International Joint Conferences on Artificial Intelligence Organization. Main track.

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*, pages 1224–1233, Austin, Texas. 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, Bonnie Lynn Webber, and Deyi Xiong. 2020. Ted-cdb: A large-scale chinese discourse relation dataset on ted talks. In *EMNLP*.

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. 2015. Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:1–1.

Jirí Mírovský, Pavlína Jínová, and Lucie Poláková. 2014. Discourse relations in the prague dependency treebank 3.0. In *COLING*.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

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 *ACL*.

Silvia Pareti and Irina Prodanof. 2010. Annotating attribution relations: Towards an Italian discourse treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *EMNLP*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77, Berlin, Germany. Association for Computational Linguistics.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *International Conference on Learning Representations*.

Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. 2020. Interactively-propagative attention learning for implicit discourse relation recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3168–3178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 799–808, Denver, Colorado. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Wei Shi and Vera Demberg. 2019a. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019b. Next sentence prediction helps implicit discourse relation classification within and across domains. 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 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Varsha Suresh and Desmond C. Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *EMNLP*.

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 (Volume 1: Long Papers)*, pages 732–742, Online. Association for Computational Linguistics.

Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2332–2342, Online. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.

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 *AAAI*.

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

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020b. Clear: Contrastive learning for sentence representation. *ArXiv*, abs/2012.15466.

Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. 2022. 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.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Seonghyeon Ye, Jiseon Kim, and Alice Oh. 2021. Efficient contrastive learning via novel data augmentation and curriculum learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1832–1838, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

Deniz Zeyrek, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5786–5798, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 PDTB Hierarchy

The hierarchies of both PDTB 2.0 and PDTB 3.0 consist of three levels, but for implicit relation recognition, so far no classification for third level labels has been done. We also focus on the hierarchy between level-1 and level 2. The PDTB-3 relation hierarchy simplifies and extends the PDTB-2 relation hierarchy. The PDTB 3.0 hierarchy not only simplifies the PDTB-2 relation hierarchy by restricting Level-3 relations to differences in directionality and eliminating rare and/or difficult-to-annotate senses, but also augments the relation hierarchy. Figure 1 and Figure 2 show PDTB 2.0 relation hierarchy and PDTB 3.0 relation hierarchy respectively.
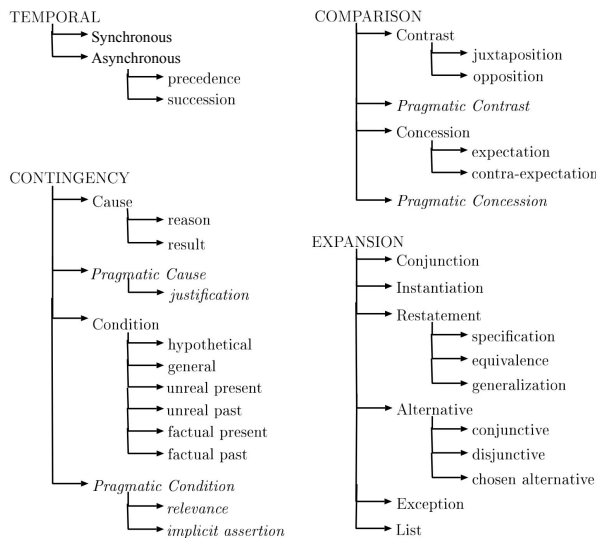
Model	Comp.	Cont	Exp.	Temp.
Liu and Li (2016)	29.15	63.33	65.10	41.03
Lan et al. (2017)	30.10	60.91	64.03	33.71
Ruan et al. (2020)	30.37	61.95	64.28	34.74
Chen et al. (2016b)	27.34	62.56	64.71	38.91
Xiang et al. (2022) (BiLSTM)	34.16	65.48	67.82	40.22
Xiang et al. (2022) (BERT)	35.83	66.77	70.00	42.13
Ours	**63.30**	**78.60**	**79.91**	**58.39**

Table 9: The results of different relations on PDTB-3 in terms of F1 (%) (top-level multi-class classification).

Second-level Label	Ours
Temp.Asynchronous	66.35
Temp.Synchrony	41.38
Cont.Cause	71.38
Cont.Cause+Belief	0.0
Cont.Condition	74.07
Cont.Purpose	96.05
Comp.Contrast	56.91
Comp.Concession	60.11
Exp.Conjunction	61.70
Exp.Equivalence	11.43
Exp.Instantiation	69.83
Exp.Level-of-detail	55.34
Exp.Manner	78.43
Exp.Substitution	63.77

Table 10: The results of different relations on PDTB-3 in terms of F1 (%) (second-level multi-class classification).

A.2 The results on relation types on PDTB-3

We also examine the classification performance on PDTB-3 in terms of Macro-F1 for the four main relation types at level-1 and 14 sense types at level-2. The results can be seen in Table 9 and Table 10. Our model has significantly better performance for all level-1 relations.

As for level-2 sense types, because there are no results of previous systems, we just show the result of 14 level-2 sense types in PDTB-3 in terms of F1.

Figure 6: The PDTB 2.0 Senses Hierarchy.

Figure 7: The PDTB 3.0 Senses Hierarchy. The leftmost column contains the Level-1 senses and the middle column, the Level-2 senses. For asymmetric relations, Level-3 senses are located in the rightmost column.