Automatic Alignment of Discourse Relations of Different Discourse Annotation Frameworks

Yingxue Fu
School of Computer Science, University of St Andrews, Scotland, UK, KY16 9SX
yf30@st-andrews.ac.uk

Abstract

Existing discourse corpora are annotated based on different frameworks, which show significant dissimilarities in definitions of arguments and relations and structural constraints. Despite surface differences, these frameworks share basic understandings of discourse relations. The relationship between these frameworks has been an open research question, especially the correlation between relation inventories utilized in different frameworks. Better understanding of this question is helpful for integrating discourse theories and enabling interoperability of discourse corpora annotated under different frameworks. However, studies that explore correlations between discourse relation inventories are hindered by different criteria of discourse segmentation, and expert knowledge and manual examination are typically needed. Some semi-automatic methods have been proposed, but they rely on corpora annotated in multiple frameworks in parallel. In this paper, we introduce a fully automatic approach to address the challenges. Specifically, we extend the label-anchored contrastive learning method introduced by Zhang et al. (2022b) to learn label embeddings during discourse relation classification. These embeddings are then utilized to map discourse relations from different frameworks. We show experimental results on RST-DT (Carlson et al., 2001) and PDTB 3.0 (Prasad et al., 2018).

Keywords: Discourse annotation, representation and processing, Discourse relations

1. Introduction

Discourse relations are an important means for achieving coherence. Previous studies have shown the benefits of incorporating discourse relations in downstream tasks, such as sentiment analysis (Wang et al., 2012), text summarization (Huang and Kurohashi, 2021) and machine comprehension (Narasimhan and Barzilay, 2015). Automatic discourse relation classification is an indispensable part of discourse parsing, which is performed under some formalisms, the notable examples including the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), based on which the RST Discourse Treebank (RST-DT) is created (Marcu, 1996), and a lexicalized Tree-Adjoining Grammar for discourse (D-LTAG) (Webber, 2004), which forms the theoretical foundation for the currently largest human-annotated discourse corpus—the Penn Discourse Treebank (PDTB) (Prasad et al., 2006, 2018)[1].

As discourse annotation has a high demand on knowledge about discourse, discourse corpora are costly to create. However, these discourse formalisms typically share similar understanding of discourse relations and their role in discourse construction. Thus, an option to enlarge discourse corpora is to align the existing discourse corpora so that they can be used jointly. This line of work starts as early as Hovy and Maier (1992), but it remains challenging to uncover the relationship between discourse relations used in different annotation frameworks.

Figure 1 shows an example of RST-style annotation. The textual spans in boxes are EDUs and the arrow-headed lines represent asymmetric discourse relations, pointing from satellites to nuclei. The labels *elab(oration)* and *attribution* denote discourse relations. As the two spans connected by the relation *same-unit* are equally salient, the relation is represented by undirected parallel lines. The spans are linked recursively until a full-coverage of the whole text is formed, as shown by the uppermost horizontal line. The vertical bars highlight the nuclei.

As RST-DT and PDTB have an overlapping section of annotated texts, the corresponding PDTB-style annotation on the same text is:

1. *the agreement "an important step forward in the strengthened debt strategy"*, **that it will "when implemented, provide significant reduction in the level of debt and debt service owed by Costa Rica."** (implicit, given, Contingency.Cause.Reason)

2. *that it will provide significant reduction in the level of debt and debt service owed by Costa Rica.*, **implemented,** (explicit, when, Temporal.Asynchronous.Succession)

3. *that it will provide significant reduction in the*

[1]We focus on RST and PDTB because our method requires a large amount of data and these two frameworks have been applied to the annotation of corpora that overlap in selected texts, thus mitigating the effect of domain shift in the results. Our method does not require corpora built on the same texts.

ISA-20 Proceedings @LREC-COLING-2024, pages 27–38

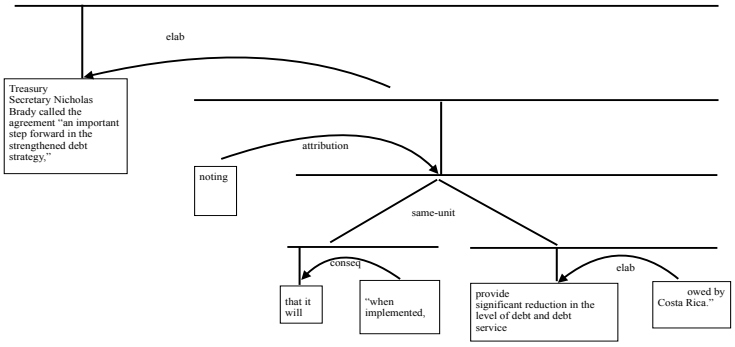


Figure 1: RST-style annotation (wsj_0624 in RST-DT).

level of debt and debt service owed by Costa Rica., **implemented,** (explicit, when, Contingency.Cause.Reason)

where Argument 1 (Arg1) is shown in italics and Argument 2 (Arg2) is in bold. The annotations in parentheses represent *relation type*, which can be implicit, explicit or others, *connective*, which is identified or inferred by annotators to signal the relation, and *sense label*, which is delimited with dots, with the first entry showing the sense label at level 1 (L1 sense), the second entry being the sense label at level 2 (L2 sense) and so on.

The task presents a challenge owing to a multitude of factors. First, different formalisms have distinctive assumptions about higher-level structures and discourse units. PDTB focuses on semantic relations between arguments, and argument identification is performed following the *Minimality Principle*, which means that only those parts that are necessary and minimally required for understanding a relation are annotated (Prasad et al., 2008). In comparison, elementary discourse units (EDUs) in RST are typically clauses. It has been shown repeatedly that segmentation criteria affect the scope of discourse relations and influence the type of relations that can be attached (Demberg et al., 2019; Benamara and Taboada, 2015; Rehbein et al., 2016).

In the first annotation of PDTB, Arg1, i.e., *the agreement "an important step forward in the strengthened debt strategy"*, is taken from the original text "Treasury Secretary Nicholas Brady called the agreement "an important step forward in the strengthened debt strategy"" and the part "Treasury Secretary Nicholas Brady called" is not covered because it does not contribute to the interpretation of the relation here. In contrast, this part is kept in an EDU in RST.

Another major difference between the two frameworks is that RST enforces a tree structure, and all the EDUs and complex discourse units (CDUs) (spans formed by adjacent elementary discourse units and adjacent lower-level spans) should be connected without crossings, while PDTB only focuses on local relations without commitment to any higher-level structure, as exemplified by the three independent annotations shown above. Previous studies (Lee et al., 2006, 2008) suggest that edge crossings and relations with shared arguments are common for PDTB. This distinction adds to the difficulty of exploring correlations of relations between the two frameworks, even if the two corpora are built on the same texts.

In addition, in RST-DT, an inventory of 78 relations is used, which can be grouped into 16 classes. These relations can be divided into *subject matter* relations (informational relations in Moore and Pollack (1992)), which are relations whose intended effect is for readers to recognize them, and *presentational* relations, which are intended to increase some inclination in readers (Mann and Thompson, 1988) (intentional relations in Moore and Pollack (1992)). For each relation, only one sense label can be attached. In contrast, PDTB adopts a three-level sense hierarchy, and more than one sense label can be annotated for a pair of arguments. As shown in the example, annotation 2 and annotation 3 are annotations for the same argument pair, but different sense labels are assigned. In previous studies that explore the alignment of RST and PDTB discourse relations, these cases typically require manual inspection to determine the closest matching PDTB relation to RST (Demberg et al., 2019). Moreover, PDTB does not take intentional relations into account but focuses on semantic and pragmatic relations.

The combination of these factors makes it challenging to investigate the relationship between discourse relations of different annotation frameworks. Even in empirical studies that make use of corpora annotated in multiple frameworks in parallel, expert knowledge and manual examination are still required. To tackle the challenge caused by differences in discourse segmentation, Demberg et al. (2019) employ the strong nuclearity hypothesis (Marcu, 2000)[2] to facilitate the string matching process of aligning PDTB arguments and RST segments. While this method alleviates the limitation of exact string matching of arguments/EDUs, it relies on a corpus annotated with multiple frameworks in parallel. Furthermore, it is conceivable that the relations left out in their analysis because of violating the principle of strong nuclearity hypothesis are not necessarily irrelevant for the goal of enabling joint usage of RST and PDTB.

In this study, we propose a fully automatic method for this task. We take inspiration from advances in label embedding techniques and an increasing body of research endeavors to harness label information in representation learning, such as supervised contrastive learning (Khosla et al., 2020; Gunel et al., 2021; Suresh and Ong, 2021). Instead of using string matching to identify the closest PDTB arguments and RST EDUs with the aim of discovering potentially analogous relations, we try to learn label embeddings of the relation inventories and compare the label embeddings.

Our contributions can be summarized as follows:

- We propose a label embedding based approach for exploring correlations between relations of different discourse annotation frameworks. The method is fully automatic and eliminates the need of matching arguments of relations.
- We conduct extensive experiments on different ways of encoding labels on RST-DT and PDTB 3.0.
- We develop a metric for evaluating the learnt label embeddings intrinsically and perform experiments to evaluate the method extrinsically.

2. Related Work

Mapping discourse relations Existing research on mapping discourse relations of different frameworks can be categorized into three types (Fu, 2022): a. identifying a set of commonly used relations across various frameworks through analysis of definitions and examples (Hovy and Maier, 1992; Bunt and Prasad, 2016; Benamara and Taboada, 2015); b. introducing a set of fundamental concepts for analyzing relations across different frameworks (Chiarcos, 2014; Sanders et al., 2018); c. mapping discourse relations directly based on corpora annotated in multiple frameworks in parallel (Rehbein et al., 2016). The third approach is closer to our method, and we summarize studies in this direction here. Rehbein et al. (2016) compare coherence relations of PDTB and CCR frameworks on the basis of a spoken corpus annotated in the two frameworks. They find that differences in annotation operationalisation and granularity of relation definition lead to many-to-many mappings. Demberg et al. (2019) show similar findings when mapping relations of RST-DT and PDTB 2.0. To mitigate issues caused by segmentation differences, they use the *strong nuclearity hypothesis* (Marcu, 2000) so that relations that have greater scope than two adjacent EDUs can be covered in their studies. With this method, Costa et al. (2023) maps RST with PDTB 3.0. Scheffler and Stede (2016) propose a method of mapping RST and PDTB relations on a German corpus annotated according to both frameworks. Explicit connectives in PDTB are used as anchors of relations, with some exceptions. It is found that 84.4% of such PDTB explicit connectives can be matched to an RST relation. The results are not surprising, as phrases that begin with a strong discourse marker are specified as EDUs (Carlson and Marcu, 2001), and a relation is likely to be attached. Stede et al. (2016) annotate a corpus with discourse information in RST and SDRT and argumentation information. A set of rules are applied to harmonize the segmentations, and structural transformation into a common dependency graph format is performed. Bourgonje and Zolotarenko (2019) try to induce PDTB implicit relations from RST annotation. Segmentation differences present a challenge, and even if the two annotations overlap in segmentation in some cases, different relations are annotated. This observation is consistent with Demberg et al. (2019).

Label embeddings Label embeddings have been proven to be useful in CV (Akata et al., 2016; Palatucci et al., 2009; Zhang et al., 2022a) and NLP taks (Wang et al., 2018; Zhang et al., 2018; Miyazaki et al., 2019). Conventionally, one-hot encoding is used to represent labels, which suffers from three problems: lack of robustness to noisy labels (Gunel et al., 2021), higher possibility of overfitting (Sun et al., 2017) and failure to capture semantic correlation of labels. Learning meaningful label representations is helpful for mitigating these problems and the semantics of labels can be used as additional information to improve model performance. It is shown that label embeddings are effective in data-imbalanced settings and zero-shot learning (Zhang et al., 2022b).

[2] A relation that holds between two spans should also hold between the nuclei of the two spans.

Label embeddings can be representations from external sources, such as BERT (Xiong et al., 2021), or can be randomly initialized (Zhang et al., 2022b). Another approach is to learn label embeddings during model training. Akata et al. (2016) propose a method of learning label embeddings from label attributes while optimizing for a classification task. Wang et al. (2018) introduce an attention mechanism that measures the compatibility of embeddings of input and labels. Additional information can be incorporated in learning label embeddings, such as label hierarchy (Chatterjee et al., 2021; Zhang et al., 2022a; Miyazaki et al., 2019) and textual description of labels (Zhang et al., 2023).

3. Method

Problem statement Given a corpus annotated in one discourse annotation framework $D_1 = \{X_m, Y_m\}_{m=1}^{M}$ and another corpus annotated in a different annotation framework $D_2 = \{X_n, Y_n\}_{n=1}^{N}$, where X denotes input sequences formed by pairs of arguments, $X_i = A_1^{(1)} ... A_a^{(1)}, A_1^{(2)} ... A_b^{(2)}$, and Y represents relation label sets of the two frameworks, $Y_{D_1} = \{y_1, y_2, ..., y_k\}$ and $Y_{D_2} = \{y_1, y_2, ..., y_c\}$. The task is to learn a correlation matrix *R* between Y_{D_1} and Y_{D_2}, which is a $2d$ matrix of shape $k \times c$. Our method is to learn embeddings for members of Y_{D_1} and Y_{D_2} and the widely used cosine similarity can be used as a measure of distance between the embedding vectors. The label embedding learning method is the same for D_1 and D_2 and we use D_1 as an example in the following.

We apply the vanilla version of label-anchored contrastive learning in Zhang et al. (2022b) as the backbone. For an input sequence X_i, we use a pre-trained language model as the input encoder f_{InEnc}. Without losing generality, we choose the popular *bert-base-uncased* model from the Huggingface transformers library (Wolf et al., 2020). For X_i pre-processed as $X_i = [CLS]\ A_1^{(1)} ... A_a^{(1)}\ [SEP]\ A_1^{(2)} ... A_b^{(2)}\ [SEP]$, the representation of the $[CLS]$ token is used as the representation of the input sequence:

$$\mathbf{E}_{X_i} = f_{InEnc}(X_i) \tag{1}$$

where the input sequence representation $\mathbf{E}_{X_i}$ is of shape $(a+b+3) \times dim$, where dim is the dimension of the output from the language model and a and b are the maximum lengths that the arguments are padded to. We empirically find that removing the non-linear transformation to $\mathbf{E}_{X_i}$ in Zhang et al. (2022b) yields better performance for our task.

We explored different options of label encoders, including: adding a BERT model (Devlin et al., 2019) ($LbEncBert$); using a RoBERTa model (Liu et al., 2020), which is trained with the next sentence prediction objective removed ($LbEncRoberta$); randomly initializing from a uniform distribution ($LbEncRand$); adding text description of the labels ($LbEncDesc$), where the label and the description are processed in the form $[CLS]label[SEP]description[SEP]$, and the representation of $[CLS]$ is used as the label representation; and adding sense hierarchy information, where we use the hierarchical contrastive loss proposed by Zhang et al. (2022a) and apply different penalty strengths to losses at different levels ($LbEncHier$). As we use language models or trainable layers as label encoders, the label embeddings are learnable.

With a label encoder g_{LbEnc}, for k total relations in D_1, we obtain a table *T* of shape $k \times lbDim$, where $lbDim$ is the output dimension of the label encoder. Thus, for a label $y_{l_{l=1}^k}$, its label embedding vector $\mathbf{E}_{y_l}$ is the $(l-1)^{\text{th}}$ row of *T*.

Instance-centered contrastive loss We apply the method in Zhang et al. (2022b) to compute the instance-centered contrastive loss $\mathcal{L}_{ICL}$:

$$\mathcal{L}_{ICL} = -\frac{1}{N} \sum_{X_i, Y_i} \log \frac{e^{\Phi(\mathbf{E}_{X_i}, \mathbf{E}_{Y_i})/\tau}}{\sum_{1 \le l \le K} e^{\Phi(\mathbf{E}_{X_i}, \mathbf{E}_{Y_l})/\tau}} \tag{2}$$

where N denotes batch size, X_i is an instance in a batch, and Y_i is its label, Φ represents a distance metric between the representations of the input and label embeddings, and cosine similarity is used in the experiment. τ denotes the temperature hyperparameter for scaling, and lower values of τ increase the influence of hard-to-separate examples in the learning process (Zhang et al., 2021). By minimizing this loss, the distance between instance representations and the corresponding class label embeddings is reduced, resulting in label embeddings that are compatible with input representations.

Label-centered contrastive loss The purpose of this loss is to reduce the distance between instances that have the same labels. For a batch with a set of unique classes C, c represents a member, P_c denotes the set of instances in a batch that have the label c and N_c represent the set of negative examples for c. A member in P_c is represented by X_p and a member in N_c is denoted by X_n. The label-centered contrastive loss $\mathcal{L}_{LCL}$ can be computed with:

$$\mathcal{L}_{LCL} = -\frac{1}{C} \sum_{c \in C} \sum_{X_p \in P_c} \log \frac{e^{\Phi(\mathbf{E}_{X_p}, \mathbf{E}_c)/\tau}}{\sum_{X_n \in N_c} e^{\Phi(\mathbf{E}_{X_n}, \mathbf{E}_c)/\tau}} \tag{3}$$

As indicated in Zhang et al. (2022b), $\mathcal{L}_{ICL}$ and $\mathcal{L}_{LCL}$ mitigate the small batch size issue encountered in other types of contrastive learning, which

makes them suitable for scenarios with limited computational resources.

We add the following two supervised losses in the training objective, which we find effective empirically.

Label-embedding-based cross-entropy loss As shown in Equation 4, a softmax function is applied to the k label embeddings in *T*, yielding a probability distribution over the k classes:

$$p(y_l) = \frac{e^{\mathbf{E}_{y_l}}}{\sum_{l=1}^{K} e^{\mathbf{E}_{y_l}}} \tag{4}$$

Let t_{y_l} denote the categorical encoding of the target y_l. The cross-entropy loss of classification based on label embeddings, denoted by $\mathcal{L}_{LEC}$, can be obtained with Equation 5:

$$\mathcal{L}_{LEC} = -\sum_{l=1}^{K} t_{y_l} \log p(y_l) \tag{5}$$

The purpose of adding this loss is to make the label embeddings better separated from each other.

Canonical multi-class cross-entropy loss We add the canonical cross-entropy loss for multi-class classification with input representations:

$$\mathcal{L}_{ICE} = -\sum_{i=1}^{N}\sum_{l=1}^{K} c_l^i \log p(c_l^i) \tag{6}$$

where N is the batch size, K is the total number of classes, and $p(c_l^i)$ is the probability predicted for a class c. With this loss, the input representations are learnt to be effective for the classification task.

The total loss is the sum of the four losses. During inference, only vector matching between the representation of an input sequence $\mathbf{E}_{X_i}$ and the k learnt embeddings $\mathbf{E}_{y_l}$ is needed, with the cosine similarity as a distance metric, for instance.

$$\hat{y} = \underset{1\leq l\leq k}{\operatorname{argmax}}(\Phi(\mathbf{E}_{X_i}, \mathbf{E}_{y_l}))) \tag{7}$$

Baseline for relation classification We run the *BertForSequenceClassification* model from the Transformers library as the baseline for discourse relation classification, which is trained with cross-entropy loss only, i.e. Equation 6.

Baseline for label embedding learning Label embeddings are generally used for improving performance in classification tasks in previous studies (Wang et al., 2018; Zhang et al., 2018; Xiong et al., 2021; Zhang et al., 2022b). To compare with a method targeted at learning good label embeddings, we implement a baseline method, which is a combination of Equation 4 and 5, but a softmax function is applied over the cosine similarities of an input $\mathbf{E}_{X_i}$ and each label embedding $\mathbf{E}_{y_l}$ in *T* here, similar to the approach adopted in Zhang et al. (2018) and Wang et al. (2018).

Metric After the model training stage, as the representations of the input sequences have been learnt for the relation classification task, we can leverage the average of the representations of input sequences X that belong to a class y_l as a proxy for the class representation, denoted by $\mathbf{H}_{y_l}$:

$$\mathbf{H}_{y_l} = \frac{1}{C}\sum_{i=1}^{C} \mathbf{E}_{X_i} \tag{8}$$

where C represents the number of instances in X.

Due to inevitable data variance, the learnt label embeddings $\mathbf{E}_{y_l}$ for a class y_l may not be the same as $\mathbf{H}_{y_l}$, but it should have a higher correlation with $\mathbf{H}_{y_l}$ than label embeddings of the other classes. Hence, we compute the correlation matrix *M* between the k learnt label embeddings $\mathbf{E}_{y_j}$ and the k class representation proxies $\mathbf{H}_{y_i}$, where $0 \leq j, i \leq k-1$, with cosine similarity as the metric of correlation:

$$M_{ij} = \Phi(\mathbf{H}_{y_i}, \mathbf{E}_{y_j}) \tag{9}$$

For each class representation proxy, we normalize its correlation scores with the k learnt label embeddings to a range of [0, 1]. The average of values at the main diagonal of *M* is adopted as an overall measure of the quality of the learnt label embeddings:

$$\mathcal{LEQ} = \frac{1}{K}\sum_{i=0}^{K-1} \widetilde{M}_{ii} \tag{10}$$

Figure 2 shows the method of intrinsic quality estimation for learnt label emebeddings.

	$\mathbf{E}_1$	$\mathbf{E}_2$	$\mathbf{E}_k$
$\mathbf{H}_1$	cos(E_1, H_1)	cos(E_2, H_1)	cos(E_k, H_1)
$\mathbf{H}_2$	cos(E_1, H_2)	cos(E_2, H_2)	cos(E_k, H_2)
$\mathbf{H}_k$	cos(E_1, H_k)	cos(E_2, H_k)	cos(E_k, H_k)

Figure 2: Illustration of the correlation matrix *M*. $\mathbf{E}_{1\ldots k}$ represents the k learnt label embeddings and $\mathbf{H}_{1\ldots k}$ denotes the k class representation proxies. After normalization, the average of the values at the diagonal (colored) is the overall measure of the quality of the learnt label embeddings.

4. Experiments

4.1. Data Preprocessing

For the purpose of our research, it would be ideal to learn label embeddings for all the relations. However, the label embeddings are trained together

with input representations in a multi-class classification task and data imbalance poses a challenge. Therefore, we focus on 16 relations for RST and PDTB L2 senses with more than 100 instances, following Kim et al. (2020).

The RST trees in RST-DT are binarized based on the procedure in Ji and Eisenstein (2014) and the spans and relations are extracted. The 78 relations are mapped to 16 classes based on the processing step in Braud et al. (2016)[3]. We take 20% from the training set of RST-DT for validation purpose.

For PDTB, we take sections 2-20 as the training set, sections 0-1 as the development set, and sections 21-22 as the test set, following Ji and Eisenstein (2015).

4.2. Hyperparameters and Training

We run each model three times with different random seeds and report the mean and standard deviation of the results. We use the AdamW optimizer (Loshchilov and Hutter, 2019) and clip L2 norm of gradients to 1.0. The learning rate is set to $1e-5$. The batch size is set to the maximum that the GPU device can accommodate. The total training epoch is set to 10 and we adopt early stop with patience of 6 on validation loss.

The temperature τ for instance-centered contrastive loss and label-centered contrastive loss is set to 0.1. For the experiment with $LbEncHier$ label encoder, the penalty factor is $2^{1/2}$ for L1 loss and 2 for L2 loss.

The learning rate for the baseline *BertForSequenceClassification* model is set to $5e-5$.

Our implementation is based on the PyTorch framework (Paszke et al., 2019) and a single 12GB RTX3060 GPU is used for all the experiments.

4.3. Results

Since we observe minimal discrepancies in data distributions between the training and test sets, we opt to utilize the test set for generating the class representation proxies necessary for the computation of the metric.

Table 1 shows the experimental results for PDTB and RST. Explicit and implicit relations for PDTB are combined. After the preprocessing step, 16 relations remain for both PDTB and RST.

It can be observed that the performance of label embedding learning on RST is lower than PDTB. Moreover, adding label embeddings generally lowers F1 compared with training with cross-entropy loss only. The decrease in F1 might be related to data sparsity when more learning objectives are added but the data amount is the same, which is visible when supplementary information of labels is added, as shown by cases of $LbEncDesc$ and $LbEncHier$. This phenomenon is rather pronounced for RST, which has a much smaller data amount. Additionally, although the label encoder $LbEncRand$ works best for the classification task, the learnt label embeddings rank the lowest among the different options. Through examination, we find that with this approach, the label embeddings of different classes are not close to the class representation proxies and we conjecture that during training, the label embeddings are mainly used as anchors, as in Zhang et al. (2022b), but the input representations are better learnt, hence the higher classification accuracy and F1 score. Zhang et al. (2022b) did not report other options of label encoders than random initialization and their focus is classification accuracy.

4.4. Data Augmentation for RST

To improve the performance on RST, we use back translation as a means of data augmentation. We translate all the files containing EDUs in the training set (only) into French and translate the French texts back into English, using Google Translate[4]. Data augmentation is not performed for *Elaboration* and *Joint*, which are the two largest classes in RST-DT, to achieve a more balanced data distribution.

Based on the results shown in Table 1, we choose $LbEncRoberta$ in the following experiments because of its good performance but results with $LbEncBert$ are comparable.

Table 2 shows the results. The F1 scores and label embedding scores are improved to a large margin. As back translation is performed at the EDU level, it is unavoidable that errors are introduced, and given that data augmentation is not performed for the two largest classes, their influence on the results is reduced, hence the lower classification accuracy.

	Acc.	F1	Label emb.
+aug.	62.75(± 0.79)	50.76(± 0.94)	92.96(± 0.90)
-aug.	65.20(± 0.07)	45.39(± 0.60)	76.56(± 0.85)

Table 2: Results for RST with data augmentation (+aug) and without data augmentation (-aug).

Figure 3 shows the T-SNE visualization plots of learnt label embeddings together with the class representation proxies for the test set of RST-DT. The label embeddings learnt with data augmentation are shown in Figure 3a in comparison with Figure 3b, where no data augmentation is performed. It is visible that in Figure 3a, more label

[3] https://bitbucket.org/chloebt/discourse/src/master/preprocess_rst/code/src/relationSet.py

[4] https://translate.google.com/

Data	Label enc.	Acc.	F1	Label emb.
PDTB total	$LbEncBert$	69.45(± 0.18)	57.80(± 0.85)	93.84(± 0.37)
	$LbEncRoberta$	69.34(± 0.46)	58.10(± 0.15)	**94.23(± 0.74)**
	$LbEncRand$	**69.87(± 0.80)**	**59.00(± 0.62)**	89.32(± 0.01)
	$LbEncDesc$	69.16(± 0.26)	57.53(± 0.14)	93.58(± 0.42)
	$LbEncHier$	69.21(± 0.45)	56.70(± 0.14)	93.67(± 0.23)
	$Baseline$	69.42(± 0.46)	58.73(± 0.78)	79.15(± 2.06)
RST	$LbEncBert$	64.62(± 0.90)	44.86(± 1.85)	**78.64(± 1.02)**
	$LbEncRoberta$	**65.20(± 0.07)**	45.39(± 0.60)	76.56(± 0.85)
	$LbEncRand$	65.09(± 0.70)	45.53(± 4.82)	69.98(± 3.10)
	$LbEncDesc$	64.62(± 0.21)	43.69(± 1.20)	74.18(± 0.91)
	$LbEncHier$	63.66(± 0.50)	41.30(± 0.39)	74.54(± 0.77)
	$Baseline$	63.55(± 0.23)	**48.57(± 0.73)**	48.21(± 1.27)

Table 1: With results over three runs collected, the Pearson correlation coefficient between classification accuracy and label embedding scores is 0.5814 and it is 0.8187 between f1 and label embedding scores, both with $p < 0.05$), which shows that the learnt label embeddings are closely related to F1 scores.

embeddings fit into the class representation proxies while in Figure 3b, label embeddings of only six classes are close to the class representation proxies, and the rest form a nebula, which suggests that the label embeddings cannot be distinguished clearly from each other. In Figure 3a, label embeddings for five relations including *Explanation*, *Textual-Organization*, *Topic-Comment*, *Evaluation* and *Topic-Change* show such behavior. *Textual-Organization*, *Topic-Comment*, and *Topic-Change* are classes with a small amount of data and it is difficult to obtain good performance on these classes in a classification task. The reason for *Explanation* and *Evaluation* is not clear, and we leave it to future work.

4.5. Separate Experiments on PDTB Explicit and Implicit Relations

Previous studies (Demberg et al., 2019; Sanders et al., 2018) indicate that it is much easier to obtain consistent results on aligning PDTB explicit relations with relations from the other frameworks, while implicit relations are generally ambiguous and the consistency is much lower. Therefore, we conducted experiments on PDTB explicit and implicit relations separately. We use $LbEncRoberta$ in the experiments. After the data preprocessing step outlined in section 4.1, 12 explicit relations and 14 implicit relations remain in the experiments.

Data	Acc.	F1	Label emb.
explicit	88.98(± 0.41)	79.19(± 0.64)	99.15(± 0.60)
implicit	56.05(± 0.56)	40.56(± 0.81)	82.21(± 0.85)

Table 3: Results of experiments on PDTB explicit relations and implicit relations.

The classification results and label embedding learning results indicate that the learnt label embeddings for PDTB explicit relations are representative of the classes while the performance on implicit relations is sub-optimal.

4.6. Ablation Study

We choose $LbEncRoberta$ and conduct ablation studies with PDTB explicit and implicit relations combined, similar to the experimental settings in Table 1. The impact of each loss can be seen in Table 4.

Loss	Acc.	F1	Label emb.
-$\mathcal{L}_{ICL}$	68.22(± 0.44)	53.65(± 1.13)	91.36(± 0.73)
-$\mathcal{L}_{LCL}$	65.02(± 0.47)	51.23(± 1.62)	80.37(± 1.42)
-$\mathcal{L}_{LEC}$	69.32(± 0.30)	57.57(± 0.87)	**94.36(± 0.37)**
-$\mathcal{L}_{ICE}$	**69.88(± 0.09)**	56.94(± 0.36)	90.79(± 0.76)
$Total$	69.34(± 0.46)	**58.10(± 0.15)**	94.23(± 0.74)

Table 4: Effect of each loss on model performance.

As shown, the label-centered contrastive loss ($\mathcal{L}_{LCL}$) is of paramount importance for the model's performance, followed by the instance-centered contrastive loss ($\mathcal{L}_{ICL}$) and canonical cross-entropy loss ($\mathcal{L}_{ICE}$). This differs from the findings in Zhang et al. (2022b), where $\mathcal{L}_{ICL}$ is the primary contributing factor to their results, indicating the distinct nature of our respective tasks. $\mathcal{L}_{LEC}$ has some effect on F1 score of the classification task.

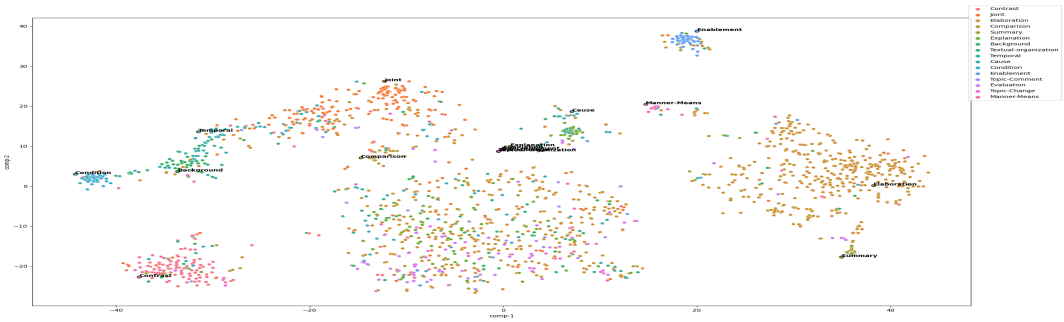
5. RST-PDTB Relation Mapping

5.1. Mapping Results

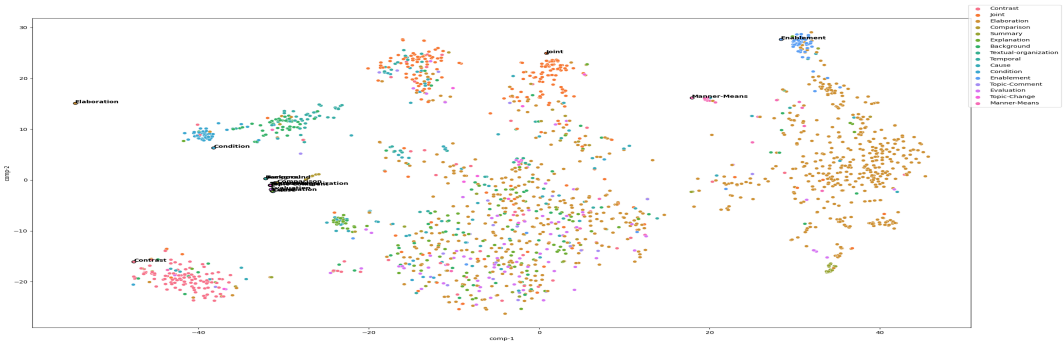
Table 5 shows the results of mapping 11 RST relations, with the five relations discussed in section 4.4 excluded, and 12 PDTB explicit relations discussed in section 4.5. Two relations with highest values in cosine similarity (greater than 0.10) are presented.

The table on the left shows the mapping results from RST's perspective. For most of the RST relations, a PDTB relation can be identified as having a much higher value ($\geq$ 0.40) than the others.

The table on the right shows the mapping results from PDTB's perspective. As relation distributions are different, it is understandable that the two perspectives are not symmetric.



(a)



(b)

Figure 3: (a) Label embeddings learnt with data augmentation. (b) Label embeddings learnt without data augmentation. For visualization, we choose the label embeddings with the highest score from the three runs.

RST	Relations in PDTB
contrast	concession(0.25), contrast(0.24)
manner-means	manner(0.30), purpose(0.25)
cause	cause(0.40), level-of-detail(0.17)
background	synchronous(0.23), manner(0.16)
condition	condition(0.39), purpose(0.18)
elaboration	concession(0.19), disjunction(0.14)
enablement	manner(0.24), purpose(0.18)
summary	contrast(0.35), level-of-detail(0.23)
joint	disjunction(0.25), synchronous(0.20)
temporal	asynchronous(0.24), purpose(0.20)
comparison	purpose(0.17), level-of-detail(0.16)

PDTB	Relations in RST
conjunction	contrast(0.22), elaboration(0.13)
concession	contrast(0.25), elaboration(0.19)
cause	cause(0.40), manner-means(0.20)
level-of-detail	manner-means(0.25), summary(0.23)
synchronous	background(0.23), joint(0.20)
disjunction	joint(0.25), temporal (0.16)
manner	manner-means(0.30), enablement(0.24)
condition	condition(0.39), summary(0.15)
substitution	manner-means(0.17), summary(0.17)
asynchronous	temporal(0.24), joint(0.19)
contrast	summary(0.35), background(0.13)
purpose	manner-means(0.25), temporal(0.20)

Table 5: Mapping between 11 RST relations and 12 PDTB explicit relations. The values in brackets represent the cosine similarity scores.

5.2. Extrinsic Evaluation

We compare our results with those provided by Costa et al. (2023), where the approach proposed in Demberg et al. (2019) is adopted but results are updated to PDTB 3.0. As shown in section 4.5, label embeddings learnt for PDTB explicit relations are more reliable and we choose to focus on the mapping between PDTB explicit relations and RST relations. Based on Table 5, we exclude PDTB's *Substitution* relation in the experiments, for which no RST relations with higher similarity are observed, and relabel 11 PDTB explicit relations with RST labels based on Table 6.

While we choose the RST label mostly based on cosine similarity shown in Table 5, we take distribution of relations into account. For example, PDTB's *Conjunction* relation is not mapped to RST's *Contrast* relation but to *Elaboration*, because *Conjunction* is a large class in PDTB, similar to *Elaboration* in RST, and relabelling in this way may keep the label distribution balanced. Meanwhile, in our preliminary experiments, mapping PDTB's *Contrast* relation to RST's *Summary* relation yields poor performance. Therefore, we relabel PDTB's *Contrast* as RST's *Contrast* relation based on the results from RST's perspective.

Similarly, we relabel PDTB explicit relations

based on the results shown in Costa et al. (2023)[5]. As their results are a mapping of 12 fine-grained RST relations and seven L2 PDTB relations, a direct mapping comparable to ours is not available. Thus, for a PDTB relation, if there are multiple mapped RST relations that fall under a broad class, the corresponding RST relation from the 16 categories is chosen, and the average of the percentages for the mapped classes is taken as the mapping strength, similar to cosine similarity in our results. For instance, PDTB *Concession* is mapped to *Contrast* (61.0%), *Antithesis* (84.0%), and *Concession* (88.0%), which are fine-grained relations under RST *Contrast*, and the mapping strength is the average of the three percentages, i.e., 0.78.

Original PDTB —Sense Labels	**RST Labels** —Our method	**RST Labels** —Costa et al. (2023)
concession	contrast (0.25)	contrast (0.78)
contrast	contrast (0.24)	contrast (0.26)
conjunction	elaboration (0.13)	joint (0.84)
manner	manner-means (0.30)	—
cause	cause (0.40)	explanation (0.69)
synchronous	background (0.23)	temporal (0.98)
condition	condition (0.39)	condition (0.84)
disjunction	joint (0.25)	—
asynchronous	temporal (0.24)	temporal (0.94)
level-of-detail	manner-means (0.25)	—
purpose	manner-means (0.25)	—

Table 6: Relabelling of PDTB explicit relations. The similarity scores are shown in brackets.

Based on our alignment results, 14964 instances of PDTB explicit relations are relabeled, and with the result in Costa et al. (2023), 13905 PDTB instances are relabeled. Adding PDTB data to RST data causes a marked performance drop. The best result is obtained with an ensemble model, which is formed by stacking a model trained with a target of minimizing supervised contrastive loss, a model trained to minimize a label embedding loss, the label embeddings being randomly initialized, and a model that takes the input for relation classification. The output distributions of the three models are averaged and used for model prediction, and a cross-entropy loss is to be reduced in addition to the supervised contrastive loss and label embedding loss. As shown in Table 7, the performance with our method is slightly higher.

	Acc.	F1
Costa et al. (2023)	62.13 $\pm$ 0.34	46.96$\pm$0.43
Our method	63.13 $\pm$ 1.12	47.95$\pm$ 1.07
-PDTB aug.	63.82$\pm$ 1.07	48.72$\pm$ 0.11

Table 7: Results of extrinsic evaluation.

6. Conclusions

We propose a method of automatically aligning discourse relations from different frameworks. By employing label embeddings that are learned concurrently with input representations during a classification task, we are able to circumvent the challenges posed by segmentation differences, a significant hurdle encountered in prior studies. We perform intrinsic and extrinsic evaluation of the results of the method. Similar to other empirical studies, our method is affected by the amount of data, and we have to exclude some relations for which there may be too little training data to learn reliable label embeddings. A comparison with a theoretical proposal, such as ISO 24617-8 (Prasad and Bunt, 2015), merits investigation in future work. The method may extend beyond labelling of discourse relations to alignment of any label sets, leaving the possibility of application to a variety of scenarios.[6]

7. Acknowledgments

We thank the anonymous reviewers for insightful feedback and suggestions. Our thanks also go to Mark-Jan Nederhof for discussions and Craig Myles for the suggestion of using the diagonal entries of the normalized correlation matrix as a metric.

8. Bibliographical References

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. 2016. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438.

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.

Peter Bourgonje and Olha Zolotarenko. 2019. Toward cross-theory discourse relation annotation. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 7–11, Minneapolis, MN. Association for Computational Linguistics.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

[5]Table 5 in their paper.

[6]We thank the anonymous reviewers for pointing out the two directions.

Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54(2001):56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Soumya Chatterjee, Ayush Maheshwari, Ganesh Ramakrishnan, and Saketha Nath Jagarlapudi. 2021. Joint learning of hyperbolic label embeddings for hierarchical multi-label classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2829–2841, Online. Association for Computational Linguistics.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. Mapping explicit and implicit discourse relations between the RST-DT and the PDTB 3.0. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 344–352, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue & Discourse*, 10(1):87–135.

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.

Yingxue Fu. 2022. Towards unification of discourse annotation frameworks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 132–142, Dublin, Ireland. 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*.

Eduard H Hovy and Elisabeth Maier. 1992. Parsimonious or profligate: how many and which discourse structure relations? Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

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.

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

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of dependencies in discourse: are dependencies in discourse more complex than in syntax? In *5th International Workshop on Treebanks and Linguistic Theories*.

Alan Lee, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2008. Departures from tree structures in discourse: Shared arguments in the penn discourse treebank. In *Proceedings of the constraints in discourse iii workshop*, pages 61–68.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1996. Building up rhetorical structure trees. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1069–1074.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT press.

Taro Miyazaki, Kiminobu Makino, Yuka Takei, Hiroki Okamoto, and Jun Goto. 2019. Label embedding using hierarchical structure of labels for Twitter classification. 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 6317–6322, Hong Kong, China. Association for Computational Linguistics.

Johanna D. Moore and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

R. Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind K. Joshi, Livio Robaldo, and Bonnie Lynn Webber. 2006. The Penn Discourse Treebank 2.0 annotation manual.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. 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).

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Ted JM Sanders, Vera Demberg, Jet Hoek, Merel CJ Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2018. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*.

Tatjana Scheffler and Manfred Stede. 2016. Mapping PDTB-style connective annotation to RST-style discourse annotation. In *Proceedings of the 13th Conference on Natural Language Processing*, pages 242–247.

Manfred Stede, Stergos Afantenos, Andreas Peldszus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *Proceedings of the Tenth International Conference on Language Resources and*

Evaluation (LREC'16), pages 1051–1058, Portorož, Slovenia. European Language Resources Association (ELRA).

Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. 2017. Label embedding network: Learning label representation for soft training of deep networks. *arXiv preprint arXiv:1710.10393*.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fei Wang, Yunfang Wu, and Likun Qiu. 2012. Exploiting discourse relations for sentiment analysis. In *Proceedings of COLING 2012: Posters*, pages 1311–1320, Mumbai, India. The COLING 2012 Organizing Committee.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.

Bonnie Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779. 2003 Rumelhart Prize Special Issue Honoring Aravind K. Joshi.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. Fusing label embedding into BERT: An efficient improvement for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750, Online. Association for Computational Linguistics.

Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. Multi-task label embedding for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4545–4553, Brussels, Belgium. Association for Computational Linguistics.

Kun Zhang, Le Wu, Guangyi Lv, Enhong Chen, Shulan Ruan, Jing Liu, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2023. Description-enhanced label embedding contrastive learning for text classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Oliver Zhang, Mike Wu, Jasmine Bayrooti, and Noah Goodman. 2021. Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022a. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669.

Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. 2022b. Label anchored contrastive learning for language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.

9. Language Resource References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. distributed via LDC. Philadelphia: Linguistic Data Consortium: LDC2002T07, Text resources, 1.0, ISLRN: 299-735-991-930-2.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. LDC. distributed via LDC. Philadelphia: Linguistic Data Consortium: LDC2019T05, Text resources, 3.0, ISLRN 977- 491-842-427-0.