Cross-Framework Generalizable Discourse Relation Classification Through Cognitive Dimensions

Yingxue Fu
School of Computer Science
University of St Andrews, Scotland, UK
fuyingxue321@gmail.com

Abstract

Existing discourse corpora annotated under different frameworks adopt distinct but somewhat related taxonomies of relations. How to integrate discourse frameworks has been an open research question. Previous studies on this topic are mainly theoretical, although such research is typically performed with the hope of benefiting computational applications. In this paper, we show how the proposal by Sanders et al. (2018) based on the Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992, 1993) can be used effectively to facilitate cross-framework discourse relation (DR) classification. To address the challenges of using predicted UDims for DR classification, we adopt the Bayesian learning framework based on Monte Carlo dropout (Gal and Ghahramani, 2016) to obtain more robust predictions. Data augmentation enabled by our proposed method yields strong performance (55.75 for RST and 55.01 for PDTB implicit DR classification in macro-averaged F1). We compare four model designs and analyze the experimental results from different perspectives. Our study shows an effective and cross-framework generalizable approach for DR classification, filling a gap in existing studies.[1]

1 Introduction

Discourse coherence relates to the way that a monologue or dialogue is organized so that it is a coherent entity, instead of a random collection of clauses or sentences. As such, coherence represents an important aspect of text quality (Webber and Joshi, 2012). Various studies have shown the benefits of incorporating discourse-level information or coherence-related training objectives in NLP tasks, such as text generation (Bosselut et al., 2018), language modelling (Iter et al., 2020; Lee et al., 2020; Stevens-Guille et al., 2022), and summarization (Xu et al., 2020).

Discourse-level analysis is typically concerned with discourse relations (Rutherford and Xue, 2015). These relations describe the links with which two segments are associated with each other and they form an integral part of discourse theories including the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and D-LTAG (Webber, 2004), which is the theoretical foundation for the PDTB framework, named after the Penn Discourse Treebank (PDTB) (Prasad et al., 2008; Webber et al., 2019). As discourse annotation is a demanding task and different discourse theories provide distinctive but often not incompatible perspectives of discourse modelling, the integration of different discourse theories has been a topic of interest for a long time (Hovy and Maier, 1992; Bunt and Prasad, 2016; Benamara and Taboada, 2015; Sanders et al., 2018; Chiarcos, 2014).

The UniDim proposal (Sanders et al., 2018), which originates from the Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992, 1993), is shown to be relatively successful in mapping PDTB and RST relations (Demberg et al., 2019). With this approach, a set of unifying dimensions (henceforth UDims) serve as *interlingua*, and relations under different frameworks can be decomposed and compared through the intermediary of it. For example, the RST relation *Contrast* can be decomposed as *negative* (at the polarity dimension, henceforth *pol*), *additive* (at the basic operation dimension, henceforth *bop*), *not applicable (NA)* (at the implication order dimension, henceforth *imp*), *objective/subjective* (at the source of coherence dimension, henceforth *soc*), and *under-specified* (at the temporality dimension, henceforth *temp*), while *Contrast* in PDTB is represented by *negative* (*pol*), *additive* (*bop*), *NA* (*imp*), *objective* (*soc*), and *under-specified* (*temp*). It is clear that the two relations are quite similar but the RST *Contrast* relation may include subjective cases (we refer those inter-

[1] Code will be released here.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 104–134

ested to Appendix A for the meaning of UDims, and Appendix B and Appendix C for a better understanding of how the relations in RST and PDTB are analyzed in terms of the UDims).

Previous studies (Roze et al., 2019; Fu, 2023; Varghese et al., 2023) demonstrate the possibility of incorporating these dimensions in discourse relation (henceforth DR) classification tasks. Varghese et al. (2023) use UDims as features for implicit DR classification, with a focus on leveraging label similarities to improve the performance of a classifier on this task. Roze et al. (2019) adopt a pipeline approach, where separate classifiers are trained for the UDims and the predicted UDims are mapped to the sense hierarchy of PDTB 2.0. As the performance on UDim classification is low, when the predicted UDims are used together to identify a sense label, the accuracy is much lower than training a classifier for DRs directly, without involving UDims. Meanwhile, the mappings are not unambiguous even between gold UDims and sense labels. The same combination of UDims can be mapped to different sense labels and the same sense labels can have different UDim representations.[2] The third challenge is that the distributions of UDims and DRs are generally imbalanced.

The study by Roze et al. (2019) shows an example of leveraging UDims in analyzing challenges of DR classification, but with their approach, UDims cannot be used effectively for DR classification tasks due to reasons discussed above. In contrast, Fu (2023) demonstrates that high performance can be achieved when gold UDims are employed for DR classification across different discourse frameworks, but the performance gains rely on availability of gold UDims, which is not feasible in realistic settings. In this study, we explore several ways of applying UDims in DR classification, and the results suggest that simply incorporating objectives of UDim classification can improve the performance on DR classification, which may be considered as empirical evidence for the correlation between UDims and DRs. However, we also notice that model performance is not a simple reflection of the relationship between UDims and DRs, for instance, a temporal relation does not necessarily have the lowest recognition accuracy when the temporal dimension is not considered in the training process, which is consistent with the findings shown in Fu (2023). In addition, we conduct experiments on using RST and PDTB data together, and the results reveal that PDTB explicit relation data is useful for data augmentation for both RST and PDTB implicit DR classification tasks.

Our contributions can be summarized as follows:

- We propose a method based on Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) to enable UDims to be applied to DR classification tasks under different frameworks, which fills a gap in existing studies (Roze et al., 2019; Fu, 2023).
- We show how UDims can be used to bridge DR classification tasks under different discourse frameworks.
- We conduct analysis of different model designs and model performance on specific relations.

2 Related Work

2.1 Cross-Framework DR Classification

Discourse connective prediction is considered a potentially effective auxiliary task for both RST DR classification (Yu et al., 2022; Yung et al., 2019) and PDTB DR classification (Qin et al., 2017; Shi and Demberg, 2019; Jiang et al., 2021; Liu and Strube, 2023). Motivated by the high performance on PDTB explicit DR classification, researchers try to convert PDTB implicit DR classification into explicit DR classification by predicting discourse connectives first. As RST does not make a clear distinction between implicit and explicit DRs in annotation, this approach is less frequently utilized for RST.

To address the challenge of limited training data for RST parsing, Braud et al. (2016) utilize multi-task learning to benefit from supervision of related tasks such as PDTB DR classification. As RST elementary discourse units (EDUs) and PDTB arguments are determined based on different criteria, they have to make adjustments to PDTB data and use sentences rather than manually annotated

[2]For example, the pattern *pos* (positive in polarity), *cau* (causal in basic operation), *NS* (under-specified in implication order), *obj* (objective in source of coherence), *NS* (under-specified in temporality), *non-specificity* (encoded by "-"), *non-alternative* ("-"), *non-conditional* ("-") and *non-goal-oriented* ("-") has two sense labels *Cause* and *Explanation* in the training set of RST, and the RST *Evaluation* relation has four patterns of UDim combinations: *pos*, *NS*, *NS*, *sub* *NS*,+, -, -, - (580 instances); *pos*, *add*, *NA*, *sub*, *NS*, +, -, -, - (272 instances); *pos*, *cau*, *bas*, *sub*, *NS*, +, -, -, - (8 instances); and *pos*, *cau*, *non-b*, *sub*, *NS*, +, -, -, - (2 instances) (see Appendix D for a full list of unique mapping patterns between UDims and DRs for RST and Appendix E for PDTB 3.0).

arguments in their experiments and ignore intra-sentential PDTB relations. Multi-task learning is also adopted in Liu et al. (2016) for PDTB implicit DR classification, where RST DR classification is treated as an auxiliary task. It shows that RST DR classification improves performance on the classification of some of PDTB Level-1 implicit DRs.

2.2 The UniDim Proposal

Under the TextLink Action, which aims at unifying existing linguistic resources on discourse structure, Sanders et al. (2018) propose a set of unifying dimensions (UDims) as an interface for different discourse frameworks to be related with each other. The UDims originate from four cognitive primitives—*basic operation*, *source of coherence*, *order of segments* (called *implication order* in Sanders et al. (2018)) and *polarity*, which are used to define coherence relations in Sanders et al. (1992), where a different approach towards representing discourse relations is taken, namely, the Cognitive approach to Coherence Relations (CCR). Compared with RST and PDTB, the CCR approach treats discourse relations as cognitive entities that can be analyzed from different dimensions, and a relation is thus described from four dimensions, such as *causal, objective, basic order, positive*, rather than with a single label, such as *Cause* in RST. Each of these dimensions functions as an attribute that has a number of possible values, for example, the *polarity* dimension allows for distinction between *positive*, *negative* or *under-specified*.

To make the taxonomy more expressive, additional dimensions are added, including *temporality*, and *specificity*, *lists* and *alternatives* for additive relations, and *conditionals* and *goal-orientedness* for causal relations. Recall that *additive* and *causal* are values under the source of coherence dimension. With these UDims, DRs from different discourse frameworks can be decomposed and compared systematically.

Demberg et al. (2019) try to validate existing proposals for mapping DRs of different discourse frameworks, and the results of their data-driven investigation exhibit higher consistency with the results obtained with the UniDim proposal, compared with the OLiA reference model (Chiarcos, 2014) and the ISO standard proposal (Bunt and Prasad, 2016).

3 Our Method

We focus on RST DR classification and PDTB implicit DR classification in this study. However, the method is generalizable, not limited to the two discourse frameworks.

3.1 UDim Extraction

Since existing discourse corpora, such as the RST Discourse Treebank (RST-DT) (Carlson et al., 2001) and PDTB, do not contain annotations of UDims, we adopt the rule-based method in Fu (2023) to obtain gold UDim values for each of the relation instances first. For RST-DT, with annotations of end labels (the original 78 relations) and nuclearity information, the mapping rules shown in Appendix B allow us to obtain UDim values. For PDTB, as the actual linear order of arguments in the original text is needed to determine values of *implication order* while the assignment of arguments does not simply follow the linear order, we first write a script to determine the linear order of arguments, and with the annotation of end labels, the mapping rules shown in Appendix C can be used to derive UDim values for each instance.

3.2 Cascaded Classifier

Given that UDims are originally intended to be used as a platform-agnostic interlingua of DRs, a natural approach is to combine all the data and train classifiers for UDims and map the predicted UDims to DRs of different discourse frameworks, based on knowledge of how UDim combinations are mapped to DRs (Appendix B and Appendix C), which is analogous to training a "universal classifier" of DRs. This approach is adopted by Roze et al. (2019), but only on PDTB 2.0, and the results are much lower than training a simple classifier for DRs directly. Moreover, Fu (2023) shows that combining PDTB and RST data does not improve the performance over using PDTB data alone for UDim classification. Therefore, instead of only using the predicted UDims for identifying DRs with a rule-based method, we deem it necessary to add DR classification as a training objective, thus forming a cascaded classifier. While this step may compromise the "universality" of the intended use of UDims, it is a necessary procedure to obtain better performance on DR classification. Figure 1 illustrates the model design.

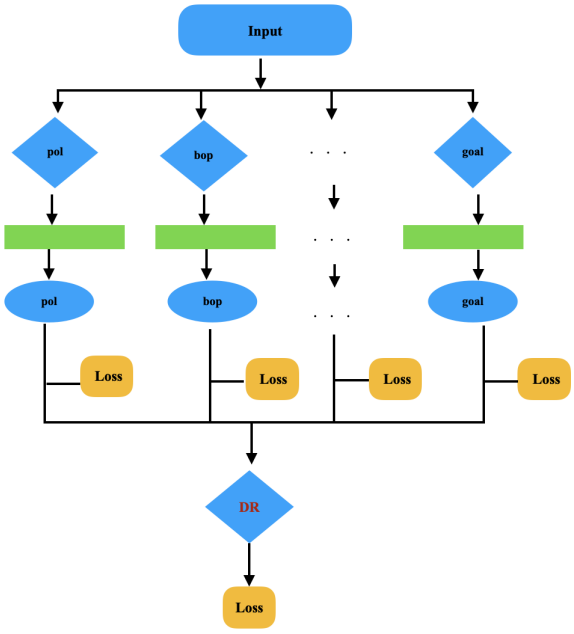

Figure 1: Cascaded classifier for DR classification. The losses in orange boxes are to be minimized. The green bars represent embedding layers. As is shown here, the input is not used for DR classification directly, which distinguishes this approach from the methods discussed in section 3.3.

3.3 Input+UDim for DR Classification

We investigate another set of methods, where the input is used for both UDim classification and DR classification. The intuition is that predicted UDims are not robust enough to be used as the only signal for DR classification, and they are better treated as attributes. Section 3.3.2 shows different model designs with this approach.

3.3.1 UDim Classification

For an input sequence X_i in a dataset with size N, i.e., $\{X_i\}_{i=1}^{N}$, X_i is formed by a pair of arguments of lengths m and n, respectively, i.e., $X_i = A_1^{(1)} \ldots A_m^{(1)}, A_1^{(2)} \ldots A_n^{(2)}$. We use a pre-trained language model as the input encoder f_{enc}. Special tokens are to be inserted based on the requirements of the chosen encoder, and X_i is typically padded to a fixed length. In our experiments, the two arguments are padded separately at the ends. The representation h of the preprocessed input sequence, denoted as $\widetilde{X_i}$, can be obtained from the encoder:

$$h = f_{enc}(\widetilde{X_i}) \tag{1}$$

A three-layer feed-forward network g, comprising a fully connected layer, a LeakyReLU activation function, followed by a dropout layer, is applied to transform h to a lower dimensional space before it is passed for UDim classification:

$$h_{UDim} = g(h) \tag{2}$$

UDims are not independent. For example, the *implication order* dimension is only applicable to causal relations, which are a category under the *basic operation* dimension. Therefore, the *basic operation* dimension functions as a parent of the *implication order* dimension. This parenthood relationship between UDims can be understood from the description of UDims in Appendix A.

Inspired by the method proposed by Gerych et al. (2021), which leverages class dependencies and conditions the prediction of child classes on the prediction of their parents, we exploit knowledge about the relationship between UDims to improve the performance on UDim classification. For instance, the embedding vector of the predicted *basic operation* dimension $E(\hat{y}_{bop})$ will be passed as features to the classification head of the *implication order* dimension, f_{impl}:

$$\tilde{y}_{impl} = softmax(f_{impl}(h_{UDim} \oplus E(\hat{y}_{bop}))) \tag{3}$$

Equation 3 shows how the prediction of the *implication order* dimension can be obtained, where $\oplus$ denotes concatenation operation.

An argmax function is required to obtain a discrete value from the predicted probability distribution, so that $E(\hat{y}_{UDim})$ can be obtained from embedding layers and passed as features for the classification of another related UDim or DR. However, this operation is non-differentiable and the training signal of one UDim cannot backpropagate to the training of the related UDims or from DRs to UDims. Therefore, we adopt the Gumbel-Softmax function (Jang et al., 2016), which is a differentiable approximation to the argmax function:

$$y_i = \frac{\exp((\log(p_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(p_j) + g_j)/\tau)} \tag{4}$$

where p_i represents a class probability for a categorical variable with k possible outcomes. $g_i \ldots g_k$ are i.i.d samples drawn from a Gumbel(0, 1) distribution, which can be sampled by drawing $\mu \sim$ Uniform(0, 1) and $g = -\log(-\log(\mu))$.

3.3.2 DR Classification

Similar to UDim classification, a three-layer feed-forward network ϕ is applied to h before it is passed

for DR classification:

$$h_{DR} = \phi(h) \tag{5}$$

We experiment with four ways of leveraging UDims in the DR classification task:

1. *TrainonGoldTestonPred*: During training, gold UDims are used and their embeddings are concatenated with h_{DR} for DR classification, so that the model learns the relationship between the input and the UDims and the target DR labels. During inference time, the embeddings of the predicted UDims are used. This is where we differ from Fu (2023), where gold UDims are still used during inference time.

2. *InputDimCat*: During both training and testing, the embeddings of predicted UDims are used, which are concatenated with h_{DR}.

3. *InputDimAtt*: During both training and testing, the embeddings of predicted UDims are combined with h_{DR} via an attention mechanism based on scaled dot product (Vaswani et al., 2017).

4. *InputForRelCls*: The hypothesis is that due to the close relationship between UDims and DRs, if the model is trained on UDim classification tasks, the performance on DR classification may be improved, even without using the predicted UDims as features, forming a scenario of multi-task learning.

Figure 2 shows the model design for *InputForRelCls*, which is also employed in the experiments on data augmentation, thus illustrated here to facilitate understanding.

Preliminary experiments show that directly using predicted UDims as features yields mixed results. This could be attributed to the utilization of *predicted* UDims, where the classification errors of these UDims might introduce noise, and combined use of these predicted UDims exacerbates uncertainty in the DR classification task. To address this challenge, the MC dropout method is employed.

3.3.3 MC Dropout

Due to the property of learning a distribution over model parameters, Bayesian networks represent a natural choice for uncertainty estimation. However, traditional Bayesian methods typically come

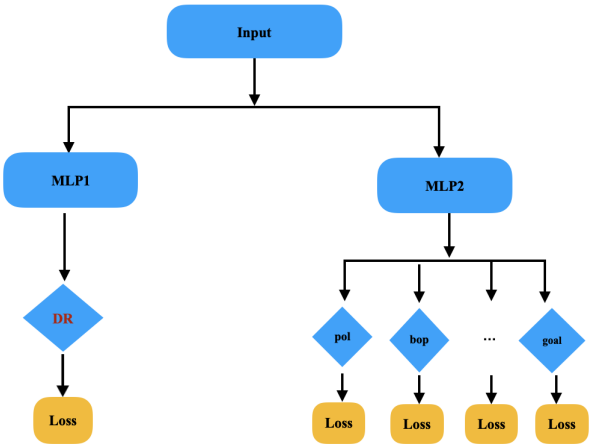

Figure 2: Model design for *InputForRelCls*. The losses in orange boxes are to be minimized.

with large computational costs, and for transformer-based models, the computational costs can be prohibitive. Gal and Ghahramani (2016) introduce the MC dropout method to tackle the challenge of uncertainty estimation in deep neural networks. Different from the standard dropout method, dropout is activated during inference time. The MC dropout method represents a lightweight Bayesian approximation.

For an input representation from the previous layer h_{i-1}, the output representation h_i of the i_{th} layer is determined with:

$$h_i = \sigma(h_{i-1}, \mathbf{W}_i, \mathbf{M}_i) \tag{6}$$

where $\mathbf{W}_i$ denotes weights of the i_{th} layer, and $\mathbf{M}_i$ is a masking matrix, with its entries being sampled from a Bernoulli distribution, and the probability of being zero is the dropout probability p. σ denotes the activation function of this layer.

For a model with l layers, the model weights ω can be expressed as a set of weight matrices for each layer: $\omega = \{\mathbf{W}_i\}_{i=1}^{l}$. With MC dropout, during inference, one can sample T sets of ω for T stochastic forward passes and the mean predicted distribution is obtained by averaging over the T passes:

$$p(y'|\mathbf{x}, \mathbf{X}, \mathbf{Y}) = \frac{1}{N}\sum_{t=1}^{T} p(y'|\mathbf{x}, [\mathbf{W}_i^t, \mathbf{M}_i^t], \ldots, [\mathbf{W}_l^t, \mathbf{M}_l^t]) \tag{7}$$

The variance can be used as an indicator of model uncertainty. As indicated by Shelmanov et al. (2021), applying MC dropout to all the

dropout layers of a transformer model yields better performance for uncertainty estimation. Even though our focus is not uncertainty estimation, the MC dropout method can be used conveniently to approximate the results of an ensemble model and we can use mean predictive distribution over multiple runs for UDim and DR classification.

3.4 Data Augmentation

Although RST and PDTB adopt different criteria for determining discourse units/arguments of DRs, data from both frameworks can be used together for UDim classification. For example, for RST DR classification, PDTB data (explicit, implicit, or both) can be used for training the model on UDim classification. Increased data amount and more diversified training data might increase model robustness in UDim classification, which may improve model performance on DR classification. Fig 3 shows a diagram of the data augmentation method.

3.5 Training

Cross-entropy loss is used for model training. Model losses for UDim classification and DR classification are added:

$$\mathcal{L}_{total} = \mathcal{L}_{UDims} + 2.0 * \mathcal{L}_{DR} \quad (8)$$

Note that there are multiple UDims involved in the experiments, even though the loss term shows them collectively as $\mathcal{L}_{UDims}$. In order to guide the model training towards DR classification, we increase the weight for DR classification loss.

4 Experiments

4.1 Data Preprocessing

The experiments on RST are carried out on RST-DT and the experiments on PDTB are performed on PDTB 3.0. As we follow the mainstream practice of preprocessing on the two corpora, the details are shown in Appendix F.

Since PDTB 3.0 contains a much larger number of articles, data amount differences between RST and PDTB may have a confounding effect in our experiments on data augmentation, because if the master task has a smaller data amount, the model may be trained to be biased towards the data of auxiliary tasks and the performance may decrease when evaluation is performed on the test set of the master task. Therefore, we try to increase the data amount for RST by back-translating data of the training set (English→French→English, translated by Google Translate), thus doubling the training data amount for RST and narrowing the data amount differences between RST, PDTB explicit relations and PDTB implicit relations. In addition, we exclude the UDim *list*, following Fu (2023), and merge sub-categories under *specificity*, making *specificity* a binary attribute, similar to *alternative*, *conditional* and *goal-orientedness*, which is also the approach adopted in Roze et al. (2019).

UDims (*abb.*)	Values	Parent
polarity(*pol*)	NS, positive, negative	-
basic operation(*bop*)	NS, additive, causal	-
source of coherence(*soc*)	NS, objective, subjective	-
implication order(*imp*)	NS, NA, basic, non-basic	bop
temporality(*temp*)	NS, anti-chronological, chronological, synchronous	-
specificity(*spec*)	specificity, non-specificity	bop
alternative(*alt*)	alternative, non-alternative	bop
conditional(*con*)	conditional, non-conditional	bop
goal-orientedness(*goal*)	goal-oriented, non-goal-oriented	bop

Table 1: UDims used in the experiments. Their abbreviations used in the paper are shown in the brackets in italics. "-" in the last column suggests that no parent passing is performed for the classification of this UDim.

Table 1 shows all the UDims used in the experiments, their abbreviations, and possible values. The parents of UDims, which are used in the method described by equation 3, are included in the last column. Appendix G shows statistics of UDims in the training sets of RST and PDTB implicit and explicit relation data, and Appendix H shows label frequency of the training sets for reference.

4.2 Implementation Details

We use the pre-trained RoBERTa$_{\text{BASE}}$ model (Liu et al., 2019) from the Transformers library (Wolf et al., 2020) as the input encoder. The embeddings of the UDims are derived from embedding layers, which are configured with learnable parameters, and the embedding vectors are initialized from uniform distributions. Hyper-parameter settings are attached in Appendix I.

Baseline The baseline is thus DR classification based on the input, involving no utilization of UDims. To ensure fair comparison, we also apply MC dropout to the baseline models, i.e., the pre-trained RoBERTa$_{\text{BASE}}$ model, and run the same

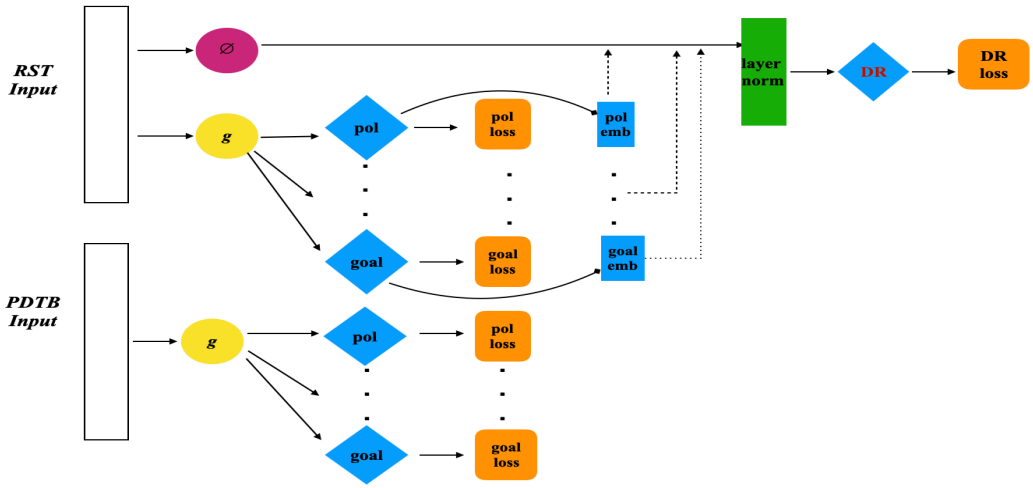

Figure 3: Data augmentation with PDTB data and the final task is RST DR classification. As we explore different ways of leveraging predicted UDims, the embeddings of the UDims are not necessarily fed as features to the DR classification task, hence represented with dashed arrow lines. The losses shown in orange boxes are to be minimized through model training for all the four methods.

number of passes to obtain the mean predictive distribution.

5 Results

We select models based on their performance measured by F1 in DR classification, and thus, they do not necessarily perform the best in terms of accuracy or in UDim classification. The Stuart-Maxwell test (Stuart, 1955; Maxwell, 1970) is used, and all the results are statistically significant (Appendix J).

5.1 DR Classification

Table 2 shows the results for RST DR classification. The best performance is achieved with *TrainonGoldTestonPred*, followed by *InputForRelCls*. In both cases, the predicted UDims are not used as features for DR classification during training. Compared with the baseline method, the models are trained for UDim classification. The results support our hypothesis that because of the association between UDims and DRs, training the model on UDim classification tasks can improve performance on DR classification.

Table 3 shows the results for PDTB implicit DR classification. A performance drop compared with the baseline is visible with the approach *TrainonGoldTestonPred*. As shown in Sanders et al. (2018,

Model	F1	Acc
Baseline	53.72	65.56
TrainonGoldTestonPred	**55.21**	66.27
InputDimCat	54.49	66.16
InputDimAtt	54.65	66.27
InputForRelCls	54.89	**66.32**

Table 2: Results for RST DR classification (The best-performing system, HITS, in DISRPT 2023 (Braud et al., 2023) achieves 50.96 in macro-averaged F1 on this corpus. As data augmentation is performed in our experiments and the preprocessing steps are different, the results are not directly comparable but shown here for reference.)

Model	F1	Acc
Baseline	52.36	60.47
TrainonGoldTestonPred	51.80	59.09
InputDimCat	52.82	**61.43**
InputDimAtt	52.93	60.67
InputForRelCls	**53.44**	60.26

Table 3: Results for PDTB implicit DR classification. Previous results on this task include 54.92 in macro-averaged F1 reported in Liu and Strube (2023), 57.62 in Long and Webber (2022) and 52.16 in Wu et al. (2023).

Model	F1	Acc
InputForRelCls	54.89	66.32
InputForRelCls+PDTBExpl	55.28	**65.72**
InputForRelCls+PDTBTotal	**55.75**	65.61
InputForRelCls+PDTBImpl	54.57	65.02

Table 4: Results for RST DR classification with data augmentation. *Baseline* refers to the approach without using UDims in training and testing in Table 2. *PDTBExpl*, *PDTBImpl* and *PDTBTotal* denote PDTB explicit relation data, implicit relation data and the combination of both parts, respectively.

Model	F1	Acc
InputForRelCls	53.44	60.26
InputForRelCls+RST	52.12	61.02
InputForRelCls+PDTBExpl	**55.01**	61.22
InputForRelCls+PDTBExpl&RST	53.05	**61.70**

Table 5: Results for PDTB implicit DR classification with data augmentation from RST data (+*RST*), from PDTB explicit relation data (+*PDTBExpl*) and from both (+*PDTBExpl&RST*).

p.52, section 5.3), implicit relations pose a challenge for the UniDim proposal, and it is likely that model performance on UDim classification is not high, when the model is trained on PDTB implicit relation data, causing a large discrepancy between training and inference time, which may result in a performance drop with *TrainonGoldTestonPred* here.

5.2 Data Augmentation

Based on the results for DR classification, we focus on the *InputForRelCls* method in this set of experiments.

Table 4 shows the results for RST DR classification under augmentation with different types of PDTB data. As is shown, data augmentation improves F1 score, but an increase in F1 does not necessarily lead to higher accuracy, which is not rare for classification on imbalanced data, suggesting that the model is trained to distinguish smaller classes. Data augmentation with total PDTB data yields the highest performance, which is expected. However, it is noticeable that adding PDTB implicit relation data causes a performance drop. This might be attributed to the high ambiguity in representing implicit relations with UDims discussed in Sanders et al. (2018).

Table 5 shows the results for PDTB implicit DR classification under augmentation with different types of data. Our method does not outperform the hierarchical sense classification method used in Long and Webber (2022) but the performance is slightly higher than that shown in Liu and Strube (2023), the best-performing method with the connective-insertion approach for converting PDTB implicit DR classification into explicit DR classification, and Wu et al. (2023), which is the SOTA performance with prompt learning.

As is shown in Table 5, adding PDTB explicit DR data is the most helpful form of data augmentation for both *InputForRelCls* and *InputDimAtt*, but adding RST data causes performance drops, possibly due to the high dissimilarity between RST data and PDTB implicit relation data.

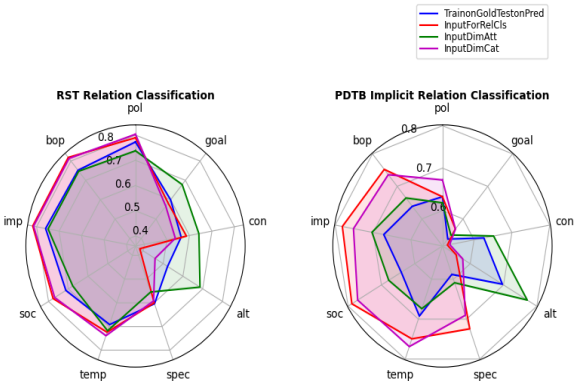

Figure 4: Correlation between DR classification loss and UDim classification losses for RST and PDTB. The abbreviations of the UDims have been explained in Table 1, and the scales represent the Pearson correlation coefficient scores. Note that the areas of different models cannot be compared between RST and PDTB, since the scales on the two plots are arranged in different ways to suit the range of the real data.

5.3 Results on Cascaded Classifier

This approach does not perform well on DR classification, but it represents a possible direction for exploration with UDims. Therefore, preliminary results are attached in Appendix K for comparison.

6 Analysis

6.1 Analysis of Different Model Designs

We examine the four approaches discussed in section 3.3.2. Losses at each training step are collected, and Pearson correlation coefficients are computed between the DR classification loss and the UDim classification losses for each model. The results are shown in Figure 4. The full results on UDim classification are shown in Appendix L for reference.

As is clear from Figure 4, for RST DR classification, the models show high correlation be-

tween DR classification and the classification of five major UDims, including *pol*, *bop*, *imp*, *soc* and *temp*, while correlation with the other UDims is not prominent. The pattern with *InputDimAtt* is different, where correlation with the UDims is basically evenly distributed, except for the smaller value at *spec*, which might be attributed to importance weighting with the attention mechanism.

For PDTB implicit DR classification, different models show divergence in their correlation strengths with different UDims. In the case of the best performing model *InputForRelCls*, the correlation with *pol* is low but the correlation with *spec* is high. We find that the performance of the model on the classification of *pol* is relatively low, and this could be a reason why the model learns to rely less on this UDim. Similar to the patterns for RST, apart from the five major UDims, the other UDims do not show high correlation with the target DR classification task, but in *TrainonGoldTestonPred* and *InputDimAtt*, relatively high correlation with *con* and *alt* in particular, is observable. The performance with *TrainonGoldTestonPred* is lower than the baseline and we can see that the total area of correlation for this model is the smallest. With *InputDimAtt*, the association area is also small, which may suggest that the attention mechanism gives more weight to h_{DR} than the predicted UDims.

6.2 Ablation Studies

Ablation studies are performed on *InputForRelCls*. Table 6 shows the UDims that, when removed, cause the lowest F1 for each DR. The full results are shown in Appendix M (RST) and Appendix N (PDTB).

RST DR	UDim
Background	*temp*
Cause	*bop*
Comparison	*spec*
Condition	*spec*
Contrast	*pol*
Elaboration	*goal*
Enablement	*alt*
Evaluation	*bop*
Explanation	*temp*
Joint	*spec*
Manner-Means	*imp*
Summary	*alt*
Temporal	*alt*
Textual-Organization	*alt*
Topic-Change	*goal*
Topic-Comment	*goal*

PDTB DR	UDim
Asynchronous	*con*
Cause	*alt*
Cause+Belief	*alt*
Concession	*goal*
Condition	*goal*
Conjunction	*soc*
Contrast	*goal*
Equivalence	*bop*
Instantiation	*pol*
Level-of-Detail	*goal*
Manner	*temp*
Purpose	*pol*
Substitution	*alt*
Synchronous	*temp*

Table 6: UDims that cause the lowest F1 for each relation. *Cause+Belief* forms a special case, where removing the UDim yields the highest performance, while removing the remaining UDims results in 0.00 for this relation.

It can be seen that the performance on some relations is consistent with the assumption about the relationship between DRs and UDims. For example, for RST, the correlation betwen *Background* and the *temp* dimension is expected. Similarly, *Cause* is indeed closely related to the *bop* dimension, which primarily distinguishes between additive and causal relations. For PDTB, the correlation between UDims and DRs is reflected in the results on *Substitution* and *Synchronous*. However, there are multiple cases when a discourse relation is not affected the most by the UDim that is supposed to be significant for it, such as RST *Condition*, which is not strongly related to *con*, but to *spec*, and *Elaboration*, which is not affected by *spec* the most, but by *goal*. Similar to what is shown in Fu (2023), model performance is not a simple reflection of the association between DRs and UDims, and it is influenced by data distributions, especially when distributions of DR and UDims are heavily imbalanced.

7 Conclusion

We propose a cross-framework generalizable approach for DR classification based on the UniDim proposal, which allows cross-framework data augmentation. With data augmentation, we obtain strong performance in macro-averaged F1 for DR classification (55.75 for RST and 55.01 for PDTB implicit DR classification). Our experiments suggest that training the model with objectives of UDim classification helps the model in DR classification, and adding PDTB explicit DR data is helpful for both RST and PDTB implicit DR classification. Our analysis shows that most of the model designs rely on five UDims, including *pol*, *bop*, *imp*, *soc* and *temp*, although differences between models are also observable. Furthermore, model performance is not a simple reflection of the expected correlation between UDims and DRs, and it is likely to be influenced by varied amounts of data for different classes. Although the present study does not involve other frameworks, such as SDRT, the proposed approach is not specific to any framework, as long as the original sense labels and rules of converting them to UDims are known.

8 Limitations

With our approach, multiple runs have to be performed during inference time, even though the number of model parameters is not increased. On the other hand, this fact justifies the choice of using

results obtained from a seemingly single run of the models, which actually involves multiple runs based on the the principle of MC dropout.

Another limitation is that in the experiments on using UDims for DR classification (without data augmentation), the improvement over the baseline is not large. However, we believe this is understandable, as more tasks are involved in the experiments (classification of nine UDims), but the data amount remains the same as the baseline, which only involves DR classification.

Compared with the upper limit of using gold UDims in DR classification, there is still a large gap. Although the UDims may be easier to understand for human annotators than the relation taxonomies employed in RST and PDTB, the performance with automatic means to predict UDims still has a large room for improvement. It remains to be tested if the difficulty in predicting these UDims forms the underlying cause for the challenges of DR classification.

We have to stress that it is beyond our scope to elaborate on the meaning of UDims and how DRs are decomposed into the combination of UDims, which falls under the CCR framework for discourse analysis. Moreover, comparison with other proposals for integrating discourse relations of different frameworks, such as the OLiA reference model and the ISO standard proposal, will be a beneficial complement to the current research. However, it is conceivable that different proposals would require different algorithmic designs to achieve good results. The current research is built on existing studies, and comparing with other proposals in computational experiments requires much more effort than the current submission can cover, and therefore, we leave it to future work.

Lastly, we are aware that discourse parsing is more than DR classification, but discourse structure is not considered in the proposed approach, similar to the focus of the work by Braud et al. (2024).

9 Ethics Statement

We do not foresee any ethical concerns with this study.

References

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.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. 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 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The 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.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

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, 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*.

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).

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.

Yingxue Fu. 2023. Discourse relations classification and cross-framework discourse relation classification through the lens of cognitive dimensions: An empirical investigation. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 21–42, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR.

Walter Gerych, Tom Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke A Rundensteiner. 2021. Recurrent bayesian classifier chains for exact multi-label classification. *Advances in Neural Information Processing Systems*, 34:15981–15992.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

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.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.

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.

Congcong Jiang, Tieyun Qian, Zhuang Chen, Kejian Tang, Shaohui Zhan, and Tao Zhan. 2021. Generating pseudo connectives with MLMs for implicit discourse relation recognition. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*, pages 113–126. Springer.

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.

Haejun Lee, Drew A. Hudson, Kangwook Lee, and Christopher D. Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562, Online. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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. *arXiv preprint arXiv:1907.11692*.

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

Ilya Loshchilov and Frank Hutter. 2018. 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.

Albert Ernest Maxwell. 1970. Comparing the classification of subjects by two independent judges. *The British Journal of Psychiatry*, 116(535):651–655.

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.

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, 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.

Charlotte Roze, Chloé Braud, and Philippe Muller. 2019. Which aspects of discourse relations are hard to learn? primitive decomposition for discourse relation classification. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 432–441, Stockholm, Sweden. Association for 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.

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*.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35.

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1993. Coherence relations in a cognitive theory of discourse representation.

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your Transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

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

Symon Stevens-Guille, Aleksandre Maskharashvili, Xintong Li, and Michael White. 2022. Generating discourse connectives with pre-trained language models: Conditioning on discourse relations helps reconstruct the PDTB. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 500–515, Edinburgh, UK. Association for Computational Linguistics.

Alan Stuart. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42(3/4):412–416.

Nobel Varghese, Frances Yung, Kaveri Anuranjana, and Vera Demberg. 2023. Exploiting knowledge about discourse relations for implicit discourse relation classification. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 99–105.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Bonnie Webber. 2004. D-ltag: extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779.

Bonnie Webber and Aravind Joshi. 2012. Discourse structure and computation: Past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

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.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage

EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics.

A UDims

Table 7 provides an overview of UDims used in the paper.

UDims (*abb.*)	Possible Values (*abb.*)	Explanations	Examples
polarity (*pol*)	positive(*pos*)	A relation is characterized by a *positive* polarity if the propositions P and Q, expressed by S_1 and S_2, respectively, have the same logical polarity and support each other.	[We like the garden]$_{S_1}$ because [it is pretty.]$_{S_2}$
	negative(*neg*)	A relation is of a *negative* polarity if the relation involves juxtaposition of $\neg P$ and P or $\neg Q$ and Q in two segments.	[The university library was closed]$_{S_1}$ although [students wanted more space for study.]$_{S_2}$
	NS	-	-
basic operation (*bop*)	additive(*add*)	If two segments are just loosely connected and only a conjunction relation $P \wedge Q$ can be inferred, the relation is additive.	[She is a painter]$_{S_1}$ and [her studio is a few blocks away.]$_{S_2}$
	causal(*cau*)	A causal relation means that two segments are strongly connected and typically, an implication relationship $P \rightarrow Q$ can be inferred.	[He immigrated to the US,]$_{S_1}$ because [his natural parents were believed to live there.]$_{S_2}$
	NS	-	-
source of coherence (*soc*)	objective(*obj*)	A relation is objective if two segments are connected because of their propositional content, and the relation holds because the connection is coherent based on world knowledge.	[It was dark outside,]$_{S_1}$ so [he lit up a candle.]$_{S_2}$
	subjective(*sub*)	A relation is subjective if the speaker's reasoning or the pragmatic effect of the relation is prominent.	[Smoking is unhealthy]$_{S_1}$ and [we should limit it.]$_{S_2}$
	NS	-	-
implication order (*imp*)	NA	This dimension distinguishes between *non-basic* and *basic* implication orders for causal relations, and it does not apply to additive relations, which are generally symmetric.	-
	basic(*bas*)	For a causal relation characterized by $P \rightarrow Q$, if S_1 expresses P and S_2 expresses Q (S_1 and S_2 are in linear order), then this relation is in basic implication order.	Because [he received a warning message,]$_{S_1}$ [he did not attend the conference.]$_{S_2}$
	non-basic(*non-b*)	In contrast to the case of basic implication order, if S_2 actually expresses P while S_1 expresses Q, this relation is in non-basic implication order.	[He did not attend the conference,]$_{S_1}$ because [he received a message telling him not to go.]$_{S_2}$
	NS	-	-
temporality (*temp*)	anti-chronological (*anti*)	If the events in two segments are not in their temporal order of occurrence, then the relation is anti-chronological.	[He went home in a low mood.]$_{S_1}$ [He had a fight with a customer and was fired.]$_{S_2}$
	chronological (*chron*)	If the events described in two segments happen in temporal order, then the relation is chronological.	[She had been stuck in a traffic jam,]$_{S_1}$ so [she was late for the opening ceremony.]$_{S_2}$
	synchronous (*sync*)	Synchronous relations are those temporal relations that feature simultaneous occurrence of events.	[The children were playing in the park]$_{S_1}$ while [their parents were chatting away.]$_{S_2}$
	NS	non-temporal relations or ambiguous cases	-

specificity (*spec*)	specificity(+)	RST and PDTB contain some relations that describe the specificity property, such as *Example*, *Definition* and *Elaboration* in RST, and *Equivalence*, *Instantiation* and *Level-of-Detail* in PDTB.	[In this light, the comparative advantages of legislative law-making become clear:]$_{S_1}$ [(1) Before it acts, the legislature typically will hear the views of representatives of all those affected by its decision, not just the immediate parties before the court; and (2) the legislature can frame "bright line" standards that create less uncertainty than the fact-bound decisions of courts.]$_{S_2}$ (wsj_2059)
	non-specificity(-)	This dimension is only applicable to additive relations. Therefore, causal relations and additive relations that do not have the property of denoting specificity are assigned the label *non-specificity*.	-
alternative (*alt*)	alternative(+)	This dimension distinguishes relations that feature two semantically alternative arguments, such as *Disjunction* in RST, and *Disjunction* and *Substitution* in PDTB.	[make their fans cheer again]$_{S_1}$ or [recapture the camaraderie of seasons past]$_{S_2}$ (wsj_0214)
	non-alternative(-)	This dimension is only applicable to additive relations. Therefore, causal relations and additive relations that do not have the property of denoting alternative propositions are assigned the label *non-alternative*.	-
conditional (*con*)	conditional(+)	Based on Sanders et al. (2018), this dimension is added to account for conditional relations, such as *Condition* in RST and PDTB, which is not possible based only on CCR dimensions (*pol*, *bop*, *soc* and *imp*).	[he will relinquish the government's so-called golden share in the company]$_{S_1}$ as long as [Jaguar shareholders agree.]$_{S_2}$ (wsj_0224)
	non-conditional(-)	This dimension is only applicable to causal relations. Therefore, additive relations and causal relations that do not have the property of being conditional are assigned the label *non-conditional*.	-
goal-orientedness (*goal*)	goal-oriented(+)	Based on Sanders et al. (2018), this dimension is added to account for relations that feature intentional and goal-oriented actions, such as *Enablement* and *Manner-Means* in RST and *Purpose* and *Manner* in PDTB.	[to clear the way]$_{S_1}$ so [the playing field is level between all contestants.]$_{S_2}$ (wsj_0224)
	non-goal-oriented(-)	This dimension is only applicable to causal relations. Therefore, additive relations and causal relations that do not have the goal-oriented property are assigned the label *non-goal-oriented*.	-

Table 7: UDims used in the experiments. Their abbreviations in this study are shown in parentheses in italics. The explanation and examples are mostly taken from Fu (2023), except for *spec*, *alt*, *con* and *goal*, which are additional dimensions deemed necessary in Sanders et al. (2018) to account for some relations in RST, PDTB and SDRT.

B Mapping Between RST Relations and UDims

Table 8 shows the mapping between RST relations and UDims, which is originally given in Sanders et al. (2018).

Class	End label	Nuc.	N-S	pol	bop	imp	soc	temp	other
Background	Background	Mono	N-S	pos/neg	add	N.A.	obj	anti/N.A.	
	Background	Mono	S-N	pos/neg	add	N.A.	obj	chron/N.A.	
	Circumstance	Mono		pos/neg	add	N.A.	obj	syn/N.A.	
Cause	Cause	Mono	N-S	pos	cau	bas	obj	chron	
	Cause	Mono	S-N	pos	cau	non-b	obj	anti	
	Cause-result	Multi		pos	cau	bas/non-b	obj	chron/anti	
	Result	Mono	N-S	pos	cau	non-b	obj	anti	
	Result	Mono	S-N	pos	cau	bas	obj	chron	
	Consequence-n	Mono	N-S	pos	cau	non-b	obj	anti	
	Consequence-n	Mono	S-N	pos	cau	bas	obj	chron	
	Consequence-s	Mono	N-S	pos	cau	bas	obj	chron	
	Consequence-s	Mono	S-N	pos	cau	non-b	obj	anti	
	Consequence	Multi		pos	cau	bas/non-b	obj	chron/anti	
Comparison	Comparison	Both		pos	add	N.A.	obj/sub	N.A.	
	Preference	Mono		neg	add	N.A.	obj/sub	N.A.	
	Analogy	Both		pos	add	N.A.	sub	N.A.	
	Proportion	Multi		pos	add/cau	any	obj/sub	any	
Conditional	Condition	Mono	N-S	pos/neg	cau	non-b	obj/sub	anti/N.A.	conditional
	Condition	Mono	S-N	pos/neg	cau	bas	obj/sub	chron/N.A.	conditional
	Hypothetical	Mono	N-S	pos	cau	non-b	sub	N.A.	conditional
	Hypothetical	Mono	S-N	pos	cau	bas	sub	N.A.	conditional
	Contingency	Mono	N-S	pos/neg	cau	non-b	obj	anti	conditional
	Contingency	Mono	S-N	pos/neg	cau	bas	obj	chron	conditional
	Otherwise	Mono	N-S	neg	cau	bas	obj/sub	chron/N.A.	conditional
	Otherwise	Multi		neg	cau	bas	obj/sub	chron/N.A.	conditional
Contrast	Contrast	Multi		neg	add	N.A.	obj/sub	any	
	Concession	Mono	N-S	neg	cau	non-b	obj/sub	anti/N.A.	
	Concession	Mono	S-N	neg	cau	bas	obj/sub	chron/N.A.	
	Antithesis	Mono		neg	add/cau	any	obj/sub	any	
Elaboration	El.-additional	Mono		pos	add	N.A.	obj/sub	N.A.	
	El.-gen.-spec.	Mono		pos	add	N.A.	obj/sub	N.A.	specificity
	El.-part-whole	Mono		pos	add	N.A.	obj	N.A.	specificity
	El.-process-step	Mono		pos	add	N.A.	obj	N.A.	specificity
	El.-object-attr.	Mono		pos	add	N.A.	obj	N.A.	specificity
	El.-set-member	Mono		pos	add	N.A.	obj	N.A.	spec.-ex.
	Example	Mono		pos	add	N.A.	obj	N.A.	spec.-ex.
	Definition	Mono		pos	add	N.A.	obj	N.A.	specificity
Enablement	Purpose	Mono	N-S	pos	cau	bas	obj/sub	chron/N.A.	goal
	Purpose	Mono	S-N	pos	cau	non-b	obj/sub	anti/N.A.	goal
	Enablement	Mono	N-S	pos	cau	non-b	obj/sub	anti/N.A.	goal
	Enablement	Mono	S-N	pos	cau	bas	obj/sub	chron/N.A.	goal
Evaluation	Evaluation	Both		pos	add/cau	any	sub	N.A.	specificity
	Interpretation	Both		pos	add/cau	any	sub	N.A.	specificity
	Conclusion	Mono	N-S	pos	cau	bas	sub	N.A.	specificity
	Conclusion	Mono	S-N	pos	cau	non-b	sub	N.A.	specificity
	Conclusion	Multi		pos	cau	bas/non-b	sub	N.A.	specificity
	Comment	Mono		pos	add	N.A.	sub	N.A.	specificity
Explanation	Evidence	Mono	N-S	pos	cau	non-b	sub	anti	
	Evidence	Mono	S-N	pos	cau	bas	sub	chron	
	Exp.-argument.	Mono	N-S	pos	cau	non-b	obj	anti	
	Exp.-argument.	Mono	S-N	pos	cau	bas	obj	chron	
	Reason	Mono	N-S	pos	cau	non-b	obj	anti	
	Reason	Mono	S-N	pos	cau	bas	obj	chron	
	Reason	Multi		pos	cau	bas/non-b	obj	chron/anti	
Joint	List	Multi		pos	add	N.A.	obj/sub	syn/chron/N.A.	list
	Disjunction	Multi		pos/neg	add	N.A.	obj/sub	syn/N.A.	alternative
Summary	Summary	Mono		pos	add	N.A.	obj	N.A.	specificity
	Restatement	Mono		pos	add	N.A.	obj	N.A.	spec.-equiv.
Temporal	Temp.-before	Mono	N-S	pos	add	N.A.	obj	chron	
	Temp.-before	Mono	S-N	pos	add	N.A.	obj	anti	
	Temp.-after	Mono	N-S	pos	add	N.A.	obj	anti	
	Temp.-after	Mono	S-N	pos	add	N.A.	obj	chron	
	Temp.-same-time	Both		pos	add	N.A.	obj	syn	
	Sequence	Multi		pos	add	N.A.	obj	chron	
	Inverted-seq.	Multi		pos	add	N.A.	obj	anti	
Manner-Means	Means	Mono	N-S	pos	cau	non-b	obj	anti	
	Means	Mono	S-N	pos	cau	bas	obj	chron	goal
Topic-Comment	Problem-sol.-n	Mono	N-S	pos	cau	non-b	obj/sub	anti/N.A.	goal
	Problem-sol.-n	Mono	S-N	pos	cau	bas	obj/sub	chron/N.A.	goal
	Problem-sol.-s	Mono	N-S	pos	cau	bas	obj/sub	chron/N.A.	goal
	Problem-sol.-s	Mono	S-N	pos	cau	non-b	obj/sub	anti/N.A.	goal
	Problem-sol.	Multi		pos	cau	bas/non-b	obj/sub	achron/anti/N.A.	goal

Table 8: Mapping between RST relations and UDims.

C Mapping Between PDTB Relations and UDims

Table 9 shows the mapping between relations in PDTB 3.0 and UDims. As the mapping given in Sanders et al. (2018) is between relations in PDTB 2.0 and UDims, we adopt the mapping table in Fu (2023).

Class_type	End label	A1-A2	pol	bop	imp	soc	temp	other
Temporal								
Synchronous			pos	add	N.A.	obj	sync	
Asynchronous	Precedence	A1-A2	pos	add	N.A.	obj	chron	
	Precedence	A2-A1	pos	add	N.A.	obj	anti	
	Succession	A1-A2	pos	add	N.A.	obj	anti	
	Succession	A2-A1	pos	add	N.A.	obj	chron	
Contingency								
Cause	Reason	A1-A2	pos	cau	non-b	obj	anti	
	Reason	A2-A1	pos	cau	bas	obj	chron	
	Result	A1-A2	pos	cau	bas	obj	chron	goal
	Result	A1-A2	pos	cau	bas	obj	chron	goal
	NegResult		neg	cau	bas	obj	chron	
Cause+Belief	Reason+Belief	A1-A2	pos	cau	non-b	sub	NS	
	Reason+Belief	A2-A1	pos	cau	bas	sub	NS	
	Result+Belief	A1-A2	pos	cau	bas	sub	NS	
	Result+Belief	A2-A1	pos	cau	non-b	sub	NS	
Cause +SpeechAct	Reason+SpeechAct	A1-A2	pos	cau	non-b	sub	NS	
	Reason+SpeechAct	A2-A1	pos	cau	bas	sub	NS	
	Result+SpeechAct	A1-A2	pos	cau	bas	sub	NS	
	Result+SpeechAct	A2-A1	pos	cau	non-b	sub	NS	
Purpose	arg1-as-goal	A1-A2	pos	cau	non-b	obj/sub	NS	goal
	arg1-as-goal	A2-A1	pos	cau	bas	obj/sub	NS	goal
	arg2-as-goal	A1-A2	pos	cau	bas	sub	NS	goal
Condition	arg1-as-cond	A1-A2	pos	cau	bas	obj/sub	NS	conditional
	arg1-as-cond	A2-A1	pos	cau	non-b	obj/sub	NS	conditional
	arg2-as-cond	A1-A2	pos	cau	non-b	obj/sub	NS	conditional
	arg2-as-cond	A2-A1	pos	cau	bas	obj/sub	NS	conditional
Condition +SpeechAct			pos	cau	bas	sub	NS	conditional
Negative -Condition	arg1-as-negcond	A1-A2	neg	cau	bas	sub	NS	conditional
	arg1-as-negcond	A2-A1	neg	cau	non-b	sub	NS	conditional
	arg2-as-negcond	A1-A2	neg	cau	non-b	sub	NS	conditional
	arg2-as-negcond	A2-A1	neg	cau	bas	sub	NS	conditional
Negative-Condition+ SpeechAct			neg	cau	bas	sub	NS	conditional
Comparison								
Concession	arg1-as-denier	A1-A2	neg	cau	non-b	obj/sub	NS	
	arg1-as-denier	A2-A1	neg	cau	bas	obj/sub	NS	
	arg2-as-denier	A1-A2	neg	cau	bas	obj/sub	NS	
	arg2-as-denier	A2-A1	neg	cau	non-b	obj/sub	NS	
Concession +SpeechAct			neg	cau	bas	sub	NS	
Contrast			neg	add	NA	obj	NS	
Similarity			pos	add	NA	obj	NS	
Expansion								
Conjunction			pos	add	NA	obj/sub	NS	
Disjunction			neg	add	NA	obj/sub	NS	alternative
Equivalence			pos	add	NA	obj/sub	NS	
Exception	arg1-as-excpt		neg	add	NA	obj/sub	NS	
	arg2-as-excpt		neg	add	NA	obj/sub	NS	
Instantiation	arg1-as-instance		pos	add	NA	obj/sub	NS	specificity
	arg2-as-instance		pos	add	NA	obj/sub	NS	specificity
Level-of-detail	arg1-as-detail		pos	add	NA	obj/sub	NS	specificity
	arg2-as-detail		pos	add	NA	obj/sub	NS	specificity
Manner	arg1-as-manner	A1-A2	pos	add	NA	obj/sub	NS	specificity
	arg2-as-manner		pos	add	NA	obj/sub	NS	specificity
Substitution	arg1-as-subst	A1-A2	neg	cau	bas	obj/sub	NS	
	arg1-as-subst	A2-A1	neg	cau	non-b	obj/sub	NS	
	arg2-as-subst	A1-A2	neg	cau	non-b	obj/sub	NS	
	arg2-as-subst	A2-A1	neg	cau	bas	obj/sub	NS	

Table 9: Mapping of sense labels of PDTB 3.0 to UniDim dimensions.

D Unique Mapping Between UDims and DRs in RST-DT

Table 10 shows unique patterns in the mapping between UDims and DRs in the training set of RST-DT. The last column shows the count of a pattern.

NS	NS	NS	NS	NS	-	-	-	-	Manner-Means	130
NS	NS	NS	NS	NS	-	-	-	-	Textual-organization	234
NS	NS	NS	NS	NS	-	-	-	-	Topic-Change	322
NS	NS	NS	NS	NS	-	-	-	-	Topic-Comment	112
NS	add	NA	NS	NS	-	+	-	-	Joint	34
NS	add	NA	obj	NS	-	-	-	-	Background	1328
NS	cau	bas	NS	NS	-	-	+	-	Condition	186
NS	cau	bas	obj	chron	-	-	+	-	Condition	14
NS	cau	non-b	NS	NS	-	-	+	-	Condition	150
NS	cau	non-b	obj	anti	-	-	+	-	Condition	26
neg	NS	NS	NS	NS	-	-	-	-	Contrast	556
neg	add	NA	NS	NS	-	-	-	-	Comparison	16
neg	add	NA	NS	NS	-	-	-	-	Contrast	588
neg	cau	bas	NS	NS	-	-	+	-	Condition	26
neg	cau	bas	NS	NS	-	-	-	-	Contrast	294
neg	cau	non-b	NS	NS	-	-	-	-	Contrast	106
pos	NS	NS	sub	NS	+	-	-	-	Evaluation	580
pos	add	NA	NS	NS	+	-	-	-	Elaboration	620
pos	add	NA	NS	NS	-	-	-	-	Comparison	368
pos	add	NA	NS	NS	-	-	-	-	Elaboration	5816
pos	add	NA	NS	NS	-	-	-	-	Joint	2898
pos	add	NA	obj	NS	+	-	-	-	Elaboration	4686
pos	add	NA	obj	NS	+	-	-	-	Summary	300
pos	add	NA	obj	anti	-	-	-	-	Temporal	124
pos	add	NA	obj	chron	-	-	-	-	Temporal	410
pos	add	NA	obj	syn	-	-	-	-	Temporal	220
pos	add	NA	sub	NS	+	-	-	-	Evaluation	272
pos	add	NA	sub	NS	+	-	-	-	Topic-Comment	4
pos	add	NA	sub	NS	-	-	-	-	Comparison	24
pos	cau	NS	NS	NS	-	-	-	-	Topic-Comment	82
pos	cau	NS	obj	NS	-	-	-	-	Cause	136
pos	cau	NS	obj	NS	-	-	-	-	Explanation	10
pos	cau	bas	NS	NS	-	-	-	+	Enablement	814
pos	cau	bas	NS	NS	-	-	-	+	Topic-Comment	14
pos	cau	bas	obj	chron	-	-	-	+	Manner-Means	18
pos	cau	bas	obj	chron	-	-	-	-	Cause	594
pos	cau	bas	obj	chron	-	-	-	-	Explanation	76
pos	cau	bas	sub	NS	+	-	-	-	Evaluation	8
pos	cau	bas	sub	NS	-	-	+	-	Condition	18
pos	cau	bas	sub	chron	-	-	-	-	Explanation	2
pos	cau	non-b	NS	NS	-	-	-	+	Enablement	76
pos	cau	non-b	obj	anti	-	-	-	-	Cause	264
pos	cau	non-b	obj	anti	-	-	-	-	Explanation	1018
pos	cau	non-b	obj	anti	-	-	-	-	Manner-Means	154
pos	cau	non-b	sub	NS	+	-	-	-	Evaluation	2
pos	cau	non-b	sub	NS	-	-	+	-	Condition	54
pos	cau	non-b	sub	anti	-	-	-	-	Explanation	278

Table 10: Unique patterns of the mapping between UDims and DRs in the training set of RST-DT.

Table 11 shows unique patterns in the mapping between UDims and DRs in the test set of RST-DT.

NS	NS	NS	NS	NS	-	-	-	-	Manner-Means	9
NS	NS	NS	NS	NS	-	-	-	-	Textual-organization	9
NS	NS	NS	NS	NS	-	-	-	-	Topic-Change	13
NS	NS	NS	NS	NS	-	-	-	-	Topic-Comment	15
NS	add	NA	NS	NS	-	+	-	-	Joint	6
NS	add	NA	obj	NS	-	-	-	-	Background	111
NS	cau	bas	NS	NS	-	-	+	-	Condition	24
NS	cau	bas	obj	chron	-	-	+	-	Condition	1
NS	cau	non-b	NS	NS	-	-	+	-	Condition	15
NS	cau	non-b	obj	anti	-	-	+	-	Condition	2
neg	NS	NS	NS	NS	-	-	-	-	Contrast	46
neg	add	NA	NS	NS	-	-	-	-	Comparison	2
neg	add	NA	NS	NS	-	-	-	-	Contrast	64
neg	cau	bas	NS	NS	-	-	-	-	Contrast	28
neg	cau	non-b	NS	NS	-	-	-	-	Contrast	8
pos	NS	NS	NS	NS	-	-	-	-	Comparison	2
pos	NS	NS	sub	NS	+	-	-	-	Evaluation	46
pos	add	NA	NS	NS	+	-	-	-	Elaboration	77
pos	add	NA	NS	NS	-	-	-	-	Comparison	24
pos	add	NA	NS	NS	-	-	-	-	Elaboration	381
pos	add	NA	NS	NS	-	-	-	-	Joint	206
pos	add	NA	obj	NS	+	-	-	-	Elaboration	338
pos	add	NA	obj	NS	+	-	-	-	Summary	32
pos	add	NA	obj	anti	-	-	-	-	Temporal	11
pos	add	NA	obj	chron	-	-	-	-	Temporal	37
pos	add	NA	obj	syn	-	-	-	-	Temporal	25
pos	add	NA	sub	NS	+	-	-	-	Evaluation	34
pos	add	NA	sub	NS	-	-	-	-	Comparison	1
pos	cau	NS	NS	NS	-	-	-	-	Topic-Comment	5
pos	cau	NS	obj	NS	-	-	-	-	Cause	13
pos	cau	NS	obj	NS	-	-	-	-	Explanation	1
pos	cau	bas	NS	NS	-	-	-	+	Enablement	43
pos	cau	bas	NS	NS	-	-	-	+	Topic-Comment	4
pos	cau	bas	obj	chron	-	-	-	+	Manner-Means	2
pos	cau	bas	obj	chron	-	-	-	-	Cause	44
pos	cau	bas	obj	chron	-	-	-	-	Explanation	2
pos	cau	non-b	NS	NS	-	-	-	+	Enablement	3
pos	cau	non-b	obj	anti	-	-	-	-	Cause	25
pos	cau	non-b	obj	anti	-	-	-	-	Explanation	95
pos	cau	non-b	obj	anti	-	-	-	-	Manner-Means	16
pos	cau	non-b	sub	NS	-	-	+	-	Condition	6
pos	cau	non-b	sub	anti	-	-	-	-	Explanation	12

Table 11: Unique patterns of UDims and DRs in the test set of RST-DT.

E Unique Mapping Between UDims and Implicit DRs in PDTB 3.0

Table 12 shows unique patterns in the mapping between UDims and implicit DRs in the training set of PDTB 3.0.

pol.	bop.	imp.	soc.	temp.	spec.	alt.	con.	goal	DR	frequency
NS	NS	NS	NS	NS	-	-	-	-	Asynchronous	12
NS	NS	NS	NS	NS	-	-	-	-	Cause	220
NS	NS	NS	NS	NS	-	-	-	-	Cause+Belief	21
NS	NS	NS	NS	NS	-	-	-	-	Concession	7
NS	NS	NS	NS	NS	-	-	-	-	Condition	14
NS	NS	NS	NS	NS	-	-	-	-	Level-of-detail	5
NS	NS	NS	NS	NS	-	-	-	-	Manner	3
NS	NS	NS	NS	NS	-	-	-	-	Purpose	453
NS	NS	NS	NS	NS	-	-	-	-	Substitution	3
neg	add	NA	obj	NS	-	-	-	-	Contrast	607
neg	cau	bas	obj/sub	NS	-	-	-	-	Concession	1123
neg	cau	non-b	obj/sub	NS	-	-	-	-	Concession	34
neg	cau	non-b	obj/sub	NS	-	-	-	-	Substitution	275
pos	add	NA	obj	anti	-	-	-	-	Asynchronous	122
pos	add	NA	obj	chron	-	-	-	-	Asynchronous	851
pos	add	NA	obj	sync	-	-	-	-	Synchronous	325
pos	add	NA	obj/sub	NS	+	-	-	-	Instantiation	1117
pos	add	NA	obj/sub	NS	+	-	-	-	Level-of-detail	2488
pos	add	NA	obj/sub	NS	+	-	-	-	Manner	188
pos	add	NA	obj/sub	NS	-	-	-	-	Conjunction	3562
pos	add	NA	obj/sub	NS	-	-	-	-	Equivalence	252
pos	cau	bas	obj	chron	-	-	-	+	Cause	2074
pos	cau	bas	obj	chron	-	-	-	-	Cause	92
pos	cau	bas	obj/sub	NS	-	-	+	-	Condition	29
pos	cau	bas	obj/sub	NS	-	-	-	+	Purpose	1
pos	cau	bas	sub	NS	-	-	-	+	Purpose	647
pos	cau	bas	sub	NS	-	-	-	-	Cause+Belief	54
pos	cau	non-b	obj	anti	-	-	-	-	Cause	2083
pos	cau	non-b	obj/sub	NS	-	-	+	-	Condition	109
pos	cau	non-b	obj/sub	NS	-	-	-	+	Purpose	1
pos	cau	non-b	sub	NS	-	-	-	-	Cause+Belief	82

Table 12: Unique patterns of UDims and implicit DRs in the training set of PDTB 3.0.

Table 13 shows unique patterns in the mapping between UDims and implicit DRs in the test set of PDTB 3.0.

Table 14 shows unique patterns in the mapping between UDims and DRs in the training set of explicit data.

pol.	bop.	imp.	soc.	temp.	spec.	alt.	con.	goal	DR	frequency
NS	NS	NS	NS	NS	-	-	-	-	Asynchronous	3
NS	NS	NS	NS	NS	-	-	-	-	Cause	20
NS	NS	NS	NS	NS	-	-	-	-	Cause+Belief	1
NS	NS	NS	NS	NS	-	-	-	-	Concession	1
NS	NS	NS	NS	NS	-	-	-	-	Condition	1
NS	NS	NS	NS	NS	-	-	-	-	Instantiation	1
NS	NS	NS	NS	NS	-	-	-	-	Manner	1
NS	NS	NS	NS	NS	-	-	-	-	Purpose	33
NS	NS	NS	NS	NS	-	-	-	-	Substitution	1
neg	add	NA	obj	NS	-	-	-	-	Contrast	53
neg	cau	bas	obj/sub	NS	-	-	-	-	Concession	88
neg	cau	non-b	obj/sub	NS	-	-	-	-	Concession	9
neg	cau	non-b	obj/sub	NS	-	-	-	-	Substitution	25
pos	add	NA	obj	anti	-	-	-	-	Asynchronous	9
pos	add	NA	obj	chron	-	-	-	-	Asynchronous	93
pos	add	NA	obj	sync	-	-	-	-	Synchronous	35
pos	add	NA	obj/sub	NS	+	-	-	-	Instantiation	123
pos	add	NA	obj/sub	NS	+	-	-	-	Level-of-detail	208
pos	add	NA	obj/sub	NS	+	-	-	-	Manner	16
pos	add	NA	obj/sub	NS	-	-	-	-	Conjunction	236
pos	add	NA	obj/sub	NS	-	-	-	-	Equivalence	30
pos	cau	bas	obj	chron	-	-	-	+	Cause	200
pos	cau	bas	obj	chron	-	-	-	-	Cause	11
pos	cau	bas	obj/sub	NS	-	-	+	-	Condition	4
pos	cau	bas	sub	NS	-	-	-	+	Purpose	56
pos	cau	bas	sub	NS	-	-	-	-	Cause+Belief	8
pos	cau	non-b	obj	anti	-	-	-	-	Cause	175
pos	cau	non-b	obj/sub	NS	-	-	+	-	Condition	10
pos	cau	non-b	sub	NS	-	-	-	-	Cause+Belief	6

Table 13: Unique patterns of UDims and implicit DRs in the test set of PDTB 3.0.

pol.	bop.	imp.	soc.	temp.	spec.	alt.	con.	goal	DR	frequency
NS	NS	NS	NS	NS -	-	-	-	Purpose	4	
neg	add	NA	obj	NS	-	-	-	-	Contrast	846
neg	add	NA	obj/sub	NS	-	+	-	-	Disjunction	228
neg	cau	bas	obj/sub	NS	-	-	-	-	Concession	3449
neg	cau	bas	obj/sub	NS	-	-	-	-	Substitution	55
neg	cau	non-b	obj/sub	NS	-	-	-	-	Concession	237
neg	cau	non-b	obj/sub	NS	-	-	-	-	Substitution	123
pos	add	NA	obj	anti	-	-	-	-	Asynchronous	737
pos	add	NA	obj	chron	-	-	-	-	Asynchronous	869
pos	add	NA	obj	sync	-	-	-	-	Synchronous	1492
pos	add	NA	obj/sub	NS	+	-	-	-	Instantiation	241
pos	add	NA	obj/sub	NS	+	-	-	-	Level-of-detail	187
pos	add	NA	obj/sub	NS	+	-	-	-	Manner	227
pos	add	NA	obj/sub	NS	-	-	-	-	Conjunction	6756
pos	cau	bas	obj	chron	-	-	-	+	Cause	374
pos	cau	bas	obj	chron	-	-	-	-	Cause	173
pos	cau	bas	obj/sub	NS	-	-	+	-	Condition	415
pos	cau	bas	obj/sub	NS	-	-	-	+	Purpose	3
pos	cau	bas	sub	NS	-	-	-	+	Purpose	202
pos	cau	non-b	obj	anti	-	-	-	-	Cause	907
pos	cau	non-b	obj/sub	NS	-	-	+	-	Condition	693
pos	cau	non-b	obj/sub	NS	-	-	-	+	Purpose	92

Table 14: Unique patterns of UDims and DRs in the training set of explicit data, shown here as supplementary material for the experiments on data augmentation.

F Preprocessing

The experiments on RST are carried out on RST-DT. We follow the gold division of the corpus for training and test sets and take 20% from the training set for validation. We utilize the preprocessing method by Ji and Eisenstein (2014) and binarize the RST trees in order to obtain pairs of discourse units linked by DRs. The 78 relations are mapped to 18 broad classes based on the template in Braud et al. (2016), but as *Same-Unit* and *Attribution* are not covered in Sanders et al. (2018), the two relations are excluded in our experiments, leaving a set of 16 RST relations.

The experiments on PDTB are performed on PDTB 3.0. We follow the data split used in Ji and Eisenstein (2015), i.e., sections 2-20 for training, sections 0-1 for validation, and sections 21-22 for testing, and discard DRs with fewer than 100 instances to alleviate data imbalance, as proposed in Kim et al. (2020), leaving 14 senses from Level-2 (L2) of the sense hierarchy.

G Statistics of UDims

We follow the format of the graph in Roze et al. (2019).

Figure 5 shows statistics of UDims for the training set of RST-DT.

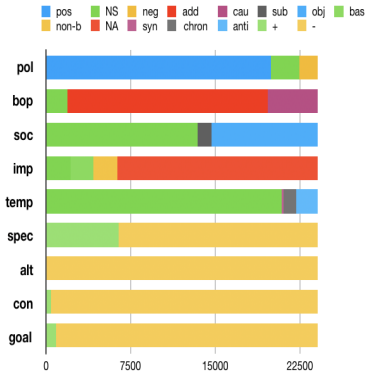

Figure 5: Statistics of UDims for the training set of RST-DT.

Figure 6 shows statistics of UDims for the training set of PDTB implicit relation data.

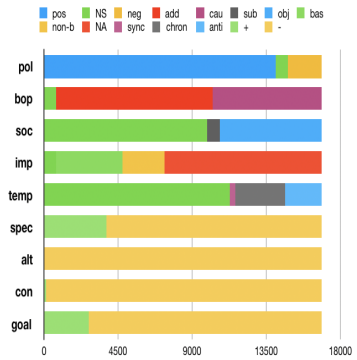

Figure 6: Statistics of UDims for the training set of PDTB implicit relation data.

H Label Frequency for Training Sets of RST and PDTB Implicit Relations

Table 15 shows label frequency in the training set of RST.

Label	Frequency
Explanation	1384
Manner-Means	302
Summary	300
Elaboration	11122
Contrast	1544
Evaluation	862
Joint	2932
Background	1328
Topic-Comment	212
Enablement	890
Cause	994
Condition	474
Topic-Change	322
Textual-Organization	234
Temporal	754
Comparison	408

Table 15: Label frequency in the training set of RST.

Table 16 shows label frequency in the training set of PDTB implicit relations.

Label	Frequency
Level-of-Detail	2493
Conjunction	3562
Concession	1164
Cause	4469
Instantiation	1117
Equivalence	252
Substitution	278
Asynchronous	985
Synchronous	325
Cause+Belief	157
Purpose	1102
Manner	191
Contrast	607
Condition	152

Table 16: Label frequency in the training set of PDTB implicit relations.

I Hyper-parameter Settings

The arguments of the input sequences are padded to a fixed length of 250 tokens, and all the model parameters are initialized with the Xavier uniform initialization (Glorot and Bengio, 2010). The output sizes of the feed-forward networks g and ϕ described in section 3.3.1 and section 3.3.2 are both set to 128 through manual tuning. The dropout probability is kept at 0.2 for all the experiments. In line with MC dropout, we keep all the dropout layers active during inference time, and run the model for UDim classification three times and obtain the average predictive distributions. The UDim embeddings are set with a dimension size of 100 in all the experiments, except for *InputDimCat* in section 3.3.2, where the dimension sizes of the UDim embeddings are set to be 2 * number of values, which we find sufficient through experimentation. Similarly, we also run the DR classifier three times and obtain the average predictive distribution. The batch size is set to the largest value that the GPU machine can accommodate.

The model learning rate is set to $1e-5$ and it is trained for a maximum of 30 epochs, with an early-stopping scheme monitoring performance improvement for DR classification on the validation set with a threshold of 7 epochs. The AdamW optimizer (Loshchilov and Hutter, 2018) is used and a warmup ratio of 0.06 is set for the scheduler. A weight decay of 0.1 is applied, and gradients are clipped to a maximum of 1.0. The implementation is based on the PyTorch machine learning framework (Paszke et al., 2019). A single A5000 GPU with a capacity of 24GB is used for all the experiments.

J Statistical Significance Test for Model Results

The saved models are used to predict DR sense labels on the test set of RST or PDTB, and different models are compared pairwise, in line with Stuart-Maxwell test.

Table 17 shows Stuart-Maxwell test of statistical significance of results for RST.

Model A	Model B	Difference	p-value
RST baseline	*InputForRelCls*	31.53487265756919	0.007444391838374066
RST baseline	*InputDimCat*	64.59896992698941	4.0166800266956324e-08
RST baseline	*InputDimAtt*	107.31845836146087	5.266424736880601e-16
RST baseline	*TrainonGoldTestonPred*	63.96165551732709	5.19121909560304e-08
InputForRelCls	*InputDimCat*	43.01215622117389	0.00015676857689742404
InputForRelCls	*InputDimAtt*	134.2049214835401	3.1837974277302182e-21
InputForRelCls	*TrainonGoldTestonPred*	57.65006061709332	6.363774121513998e-07
InputDimCat	*InputDimAtt*	178.08228924962432	5.7895617480969986e-30
InputDimCat	*TrainonGoldTestonPred*	75.50117669436882	4.596013953778202e-10
InputDimAtt	*TrainonGoldTestonPred*	90.04348995437392	9.735088746634535e-13

Table 17: Statistical significance test for results of RST models. Compared with the baseline, all the model results are statistically significant.

Table 18 shows Stuart-Maxwell test of statistical significance of results for PDTB implicit DR classification.

Model A	Model B	Difference	p-value
PDTB impl baseline	*InputForRelCls*	96.37746867104653	8.322748692517058e-15
PDTB impl baseline	*InputDimCat*	97.67850546549909	4.6673553418486555e-15
PDTB impl baseline	*InputDimAtt*	35.46139837486632	0.0007189097931712012
PDTB impl baseline	*TrainonGoldTestonPred*	106.07628543197474	1.0923958667709962e-16
InputForRelCls	*InputDimCat*	98.47524634808387	3.2737145426570587e-15
InputForRelCls	*InputDimAtt*	154.30765485412508	2.788891357957465e-26
InputForRelCls	*TrainonGoldTestonPred* 95.28342048662472	1.3526258777197362e-14	
InputDimCat	*InputDimAtt*	108.8129041233804	3.1898108418072124e-17
InputDimCat	*TrainonGoldTestonPred*	155.30024500712676	1.7579323354258866e-26
InputDimAtt	*TrainonGoldTestonPred*	136.33343659405318	1.1401098443142066e-22

Table 18: Statistical significance test for results of PDTB models. As is shown, all the model results are statistically significant.

K Detailed Results for Cascaded Classifier

This part involves two experiments. To get an estimate of the upper limit of only using UDims for DR classification, we experiment with using a single MLP to predict DRs based on gold UDims (*MLPGoldUDims*), which represents an upper limit of the approach of *universal classifier*.

The second experiment simulates the cascaded classifier, where UDim classification is performed first and predicted UDims are used for DR classification. In this experiment, input is only used for UDim classification and the predicted UDims are combined via an attention mechanism for DR classification. The training objective is to minimize losses of DR classification and UDim classification.

Table 19 shows the results, indicating a large gap between using predicted UDims (*UDimAtt*, i.e. using attention of embeddings of predicted UDims) and gold UDims (*MLPGoldUDims*) for DR classification.

	F1	Acc.
RST(*MLPGoldUDims*)	56.55	79.60
RST(*UDimAtt*)	32.24	55.50
PDTB impl.((*MLPGoldUDims*)	73.00	85.79
PDTB impl.(*UDimAtt*)	41.69	56.21

Table 19: Results on cascaded classification of UDims and DRs. *UDimAtt* denotes combining predicted UDims with an attention mechanism, which performs better than simple concatenation of predicted UDims here.

Table 20 shows the detailed performance on DR classification for RST, based on the cascaded classifier.

	precision	recall	f1	frequency
Background	44.44	36.04	39.80	111
Cause	36.36	24.39	29.20	82
Comparison	0.00	0.00	0.00	29
Condition	76.09	72.92	74.47	48
Contrast	68.32	75.34	71.66	146
Elaboration	59.42	82.79	69.19	796
Enablement	62.96	73.91	68.00	46
Evaluation	30.14	27.50	28.76	80
Explanation	37.42	55.45	44.69	110
Joint	37.50	1.42	2.73	212
Manner-Means	0.00	0.00	0.00	27
Summary	0.00	0.00	0.00	32
Temporal	65.96	42.47	51.67	73
Textual-Organization	33.33	11.11	16.67	9
Topic-Change	13.79	30.77	19.05	13
Topic-Comment	0.00	0.00	0.00	24

Table 20: Detailed results of RST DR classification with the cascaded classifier.

Table 21 shows test performance on UDim classification for RST, with the cascaded classifier based on *UDimAtt*.

Model	pol F1	bop F1	impl F1	soc F1	temp F1	spec F1	alt F1	con F1	goal F1	pol acc	bop acc	impl acc	soc acc	temp acc	spec acc	alt acc	con acc	goal acc
UDimAtt	74.66	58.44	55.41	63.02	48.16	83.09	83.29	88.85	84.52	88.36	78.56	77.15	74.48	86.89	86.13	99.84	98.91	98.31

Table 21: Results for UDim classification on RST for the cascaded classifier based on *UDimAtt*

Table 22 shows the detailed performance on DR classification for PDTB implicit relation data, based on the cascaded classifier.

Table 23 shows test performance on UDim classification for PDTB implicit relation data, with the cascaded classifier based on *UDimAtt*.

L Full Results for UDim Classification on RST and PDTB

Table 24 shows test performance on UDim classification for RST.

Table 25 shows test performance on UDim classification for PDTB implicit DRs.

	precision	recall	f1	frequency
Asynchronous	65.00	74.29	69.33	105
Cause	71.71	63.05	67.10	406
Cause+Belief	00.00	00.00	00.00	15
Concession	59.38	58.16	58.76	98
Condition	77.78	46.67	58.33	15
Conjunction	47.63	68.22	56.10	236
Contrast	55.36	58.49	56.88	53
Equivalence	00.00	00.00	00.00	30
Instantiation	00.00	00.00	00.00	124
Level-of-detail	40.00	58.65	47.56	208
Manner	00.00	00.00	00.00	17
Purpose	79.05	93.26	85.57	89
Substitution	42.11	61.54	50.00	26
Synchronous	66.67	22.86	34.04	35

Table 22: Detailed results of PDTB implicit DR classification with the cascaded classifier.

Model	pol F1	bop F1	impl F1	soc F1	temp F1	spec F1	alt F1	con F1	goal F1	pol acc	bop acc	impl acc	soc acc	temp acc	spec acc	alt acc	con acc	goal acc
UDimAtt	68.62	68.97	66.46	71.28	64.66	82.55	100.00	84.66	80.80	87.51	78.65	76.32	76.87	80.44	87.58	100.00	99.52	89.29

Table 23: Results for UDim classification on PDTB implicit relation data for the cascaded classifier based on *UDimAtt*

Model	pol F1	bop F1	impl F1	soc F1	temp F1	spec F1	alt F1	con F1	goal F1	pol acc	bop acc	impl acc	soc acc	temp acc	spec acc	alt acc	con acc	goal acc
TrainonGold TestonPred	73.65	57.78	57.06	60.69	47.18	82.50	64.22	87.82	85.16	87.43	78.45	77.53	73.78	88.79	85.91	99.73	98.91	98.37
InputDimCat	75.33	57.72	56.88	62.57	48.43	82.64	79.95	89.30	83.52	88.41	78.13	77.64	74.48	87.49	86.02	99.78	98.97	98.20
InputDimAtt	74.09	59.02	56.32	60.85	45.42	82.72	74.95	88.42	86.15	87.60	77.86	76.71	75.46	87.21	86.40	99.78	98.86	98.48
InputFor RelCls	73.19	60.34	58.33	61.39	46.41	82.53	83.29	88.36	85.16	87.54	78.84	78.13	75.14	87.00	86.45	99.84	98.91	98.37

Table 24: Results for UDim classification on RST.

Model	pol F1	bop F1	impl F1	soc F1	temp F1	spec F1	alt F1	con F1	goal F1	pol acc	bop acc	impl acc	soc acc	temp acc	spec acc	alt acc	con acc	goal acc
TrainonGold TestonPred	66.84	66.41	63.50	69.61	58.10	81.42	100.00	84.66	77.10	86.55	75.77	72.82	74.95	76.53	87.17	100.00	99.52	87.17
InputDimCat	68.59	69.71	66.43	72.12	60.72	82.46	100.00	81.68	79.84	87.71	77.97	74.74	76.53	79.75	87.37	100.00	99.45	88.81
InputDimAtt	69.03	68.40	65.66	74.24	59.93	80.16	100.00	77.10	81.08	88.95	77.49	74.19	77.08	77.21	85.86	100.00	99.31	89.09
InputFor RelCls	66.27	65.25	62.48	70.66	58.76	82.10	100.00	84.66	78.92	86.41	75.77	73.03	75.22	77.97	87.71	100.00	99.52	88.81

Table 25: Results for UDim classification on PDTB implicit relation data.

M Ablation Studies for RST

Table 26 presents results of ablation studies for RST DR classification.

DR	UDim	P.	R.	F1
Background	*-pol*	44.55	44.14	44.34
	-bop	48.28	37.84	42.42
	-imp	48.89	39.64	43.78
	-soc	44.55	40.54	42.45
	-temp	45.65	37.84	41.38
	-spec	45.26	38.74	41.75
	-alt	43.56	39.64	41.51
	-con	41.07	41.44	41.26
	-goal	42.86	43.24	43.05
Cause	*-pol*	42.86	21.95	29.03
	-bop	33.33	25.61	28.97
	-imp	34.92	26.83	30.34
	-soc	32.00	29.27	30.57
	-temp	35.82	29.27	32.21
	-spec	38.71	29.27	33.33
	-alt	34.25	30.49	32.26
	-con	36.67	26.83	30.99
	-goal	35.29	29.27	32.00
Comparison	*-pol*	52.00	44.83	48.15
	-bop	54.17	44.83	49.06
	-imp	51.61	55.17	53.33
	-soc	66.67	41.38	51.06
	-temp	52.17	41.38	46.15
	-spec	47.62	34.48	40.00
	-alt	50.00	41.38	45.28
	-con	58.33	48.28	52.83
	-goal	46.43	44.83	45.61
Condition	*-pol*	82.50	68.75	75.00
	-bop	87.50	72.92	79.55
	-imp	80.00	75.00	77.42
	-soc	80.43	77.08	78.72
	-temp	73.47	75.00	74.23
	-spec	77.27	70.83	73.91
	-alt	83.72	75.00	79.12
	-con	80.95	70.83	75.56
	-goal	76.09	72.92	74.47
Contrast	*-pol*	68.75	67.81	68.28
	-bop	75.89	73.29	74.56
	-imp	78.57	67.81	72.79
	-soc	73.76	71.23	72.47
	-temp	73.15	74.66	73.90
	-spec	75.19	68.49	71.68
	-alt	75.36	71.23	73.24
	-con	72.41	71.92	72.16
	-goal	78.79	71.23	74.82
Elaboration	*-pol*	71.61	83.67	77.17
	-bop	71.84	83.67	77.31
	-imp	72.23	82.04	76.82
	-soc	70.75	84.80	77.14
	-temp	73.66	81.16	77.23
	-spec	73.77	81.28	77.35
	-alt	73.36	84.05	78.34
	-con	76.74	81.66	79.12
	-goal	74.91	78.77	76.79
Enablement	*-pol*	71.43	76.09	73.68
	-bop	77.27	73.91	75.56
	-imp	72.00	78.26	75.00
	-soc	73.91	73.91	73.91
	-temp	68.63	76.09	72.16
	-spec	64.41	82.61	72.38
	-alt	66.67	78.26	72.00
	-con	75.00	78.26	76.60
	-goal	74.00	80.43	77.08
Evaluation	*-pol*	39.68	31.25	34.97
	-bop	38.89	26.25	31.34
	-imp	41.67	31.25	35.71
	-soc	36.62	32.50	34.44
	-temp	42.00	26.25	32.31
	-spec	36.84	35.00	35.90
	-alt	44.90	27.50	34.11
	-con	40.58	35.00	37.58
	-goal	33.33	35.00	34.15
Explanation	*-pol*	50.55	41.82	45.77
	-bop	46.85	47.27	47.06
	-imp	51.16	40.00	44.90
	-soc	48.19	36.36	41.45
	-temp	43.30	38.18	40.58
	-spec	54.00	49.09	51.43
	-alt	50.00	44.55	47.12
	-con	49.49	44.55	46.89
	-goal	38.69	48.18	42.91

Joint	*-pol*	67.10	73.11	69.98
	-bop	69.12	70.75	69.93
	-imp	65.97	74.06	69.78
	-soc	70.98	64.62	67.65
	-temp	63.60	75.00	68.83
	-spec	63.60	71.70	67.41
	-alt	70.51	72.17	71.33
	-con	64.43	76.89	70.11
	-goal	67.26	71.70	69.41
Manner-Means	*-pol*	75.00	44.44	55.81
	-bop	75.00	44.44	55.81
	-imp	71.43	37.04	48.78
	-soc	68.75	40.74	51.16
	-temp	66.67	44.44	53.33
	-spec	73.68	51.85	60.87
	-alt	65.00	48.15	55.32
	-con	57.14	44.44	50.00
	-goal	72.22	48.15	57.78
Summary	*-pol*	65.00	40.62	50.00
	-bop	66.67	43.75	52.83
	-imp	80.00	50.00	61.54
	-soc	80.00	37.50	51.06
	-temp	85.71	37.50	52.17
	-spec	75.00	46.88	57.69
	-alt	61.90	40.62	49.06
	-con	68.18	46.88	55.56
	-goal	71.43	46.88	56.60
Temporal	*-pol*	65.22	41.10	50.42
	-bop	66.04	47.95	55.56
	-imp	54.29	52.05	53.15
	-soc	66.67	41.10	50.85
	-temp	54.69	47.95	51.09
	-spec	68.52	50.68	58.27
	-alt	59.18	39.73	47.54
	-con	61.22	41.10	49.18
	-goal	72.97	36.99	49.09
Textual-Organization	*-pol*	66.67	88.89	76.19
	-bop	63.64	77.78	70.00
	-imp	72.73	88.89	80.00
	-soc	66.67	88.89	76.19
	-temp	72.73	88.89	80.00
	-spec	66.67	88.89	76.19
	-alt	57.14	88.89	69.57
	-con	66.67	88.89	76.19
	-goal	66.67	88.89	76.19
Topic-Change	*-pol*	62.50	38.46	47.62
	-bop	46.15	46.15	46.15
	-imp	38.46	38.46	38.46
	-soc	57.14	30.77	40.00
	-temp	45.45	38.46	41.67
	-spec	41.67	38.46	40.00
	-alt	62.50	38.46	47.62
	-con	50.00	38.46	43.48
	-goal	36.36	30.77	33.33
Topic-Comment	*-pol*	45.45	20.83	28.57
	-bop	37.50	25.00	30.00
	-imp	41.18	29.17	34.15
	-soc	46.67	29.17	35.90
	-temp	40.00	25.00	30.77
	-spec	46.67	29.17	35.90
	-alt	52.94	37.50	43.90
	-con	40.00	41.67	40.82
	-goal	31.58	25.00	27.91

Table 26: Ablation studies for RST, based on *InputForRelCls*. The lowest F1 scores are shown in blue, although there are cases when the differences between values are quite small.

N Ablation Studies for PDTB Implicit DR Classification

Table 27 presents results of ablation studies for PDTB implicit DR classification.

DR	UDim	P.	R.	F1
Asynchronous	*-pol*	63.11	61.90	62.50
	-bop	74.73	64.76	69.39
	-imp	66.04	66.67	66.35
	-soc	56.15	69.52	62.13
	-temp	68.09	60.95	64.32
	-spec	65.35	62.86	64.08
	-alt	63.55	64.76	64.15
	-con	59.65	64.76	62.10
	-goal	64.15	64.76	64.45
Cause	*-pol*	66.58	66.26	66.42
	-bop	64.99	69.95	67.38
	-imp	64.93	67.49	66.18
	-soc	64.49	60.84	62.61
	-temp	65.30	66.75	66.02
	-spec	69.83	59.85	64.46
	-alt	69.28	56.65	62.33
	-con	65.26	64.78	65.02
	-goal	71.64	59.11	64.78
Cause+Belief	*-pol*	11.11	06.67	08.33
	-bop	00.00	00.00	00.00
	-imp	00.00	00.00	00.00
	-soc	09.09	06.67	07.69
	-temp	00.00	00.00	00.00
	-spec	00.00	00.00	00.00
	-alt	10.71	20.00	13.95
	-con	00.00	00.00	00.00
	-goal	20.00	06.67	10.00
Concession	*-pol*	58.54	48.98	53.33
	-bop	63.41	53.06	57.78
	-imp	58.43	53.06	55.61
	-soc	66.22	50.00	56.98
	-temp	57.14	53.06	55.03
	-spec	50.00	62.24	55.45
	-alt	55.45	57.14	56.28
	-con	70.77	46.94	56.44
	-goal	59.15	42.86	49.70
Condition	*-pol*	77.78	46.67	58.33
	-bop	81.82	60.00	69.23
	-imp	80.00	53.33	64.00
	-soc	81.82	60.00	69.23
	-temp	83.33	66.67	74.07
	-spec	77.78	46.67	58.33
	-alt	87.50	46.67	60.87
	-con	77.78	46.67	58.33
	-goal	71.43	33.33	45.45
Conjunction	*-pol*	56.68	66.53	61.21
	-bop	53.31	71.61	61.12
	-imp	54.58	63.14	58.55
	-soc	50.48	67.37	57.71
	-temp	55.16	65.68	59.96
	-spec	53.77	69.49	60.63
	-alt	50.31	69.49	58.36
	-con	51.44	68.22	58.65
	-goal	49.46	77.54	60.40
Contrast	*-pol*	61.11	41.51	49.44
	-bop	43.64	45.28	44.44
	-imp	48.84	39.62	43.75
	-soc	55.56	47.17	51.02
	-temp	60.53	43.40	50.55
	-spec	52.17	45.28	48.48
	-alt	52.50	39.62	45.16
	-con	45.83	41.51	43.56
	-goal	42.59	43.40	42.99
Equivalence	*-pol*	29.73	36.67	32.84
	-bop	33.33	03.33	06.06
	-imp	25.93	23.33	24.56
	-soc	18.42	23.33	20.59
	-temp	28.00	23.33	25.45
	-spec	22.22	13.33	16.67
	-alt	40.00	20.00	26.67
	-con	18.52	33.33	23.81
	-goal	29.73	36.67	32.84
Instantiation	*-pol*	80.25	52.42	63.41
	-bop	71.15	59.68	64.91
	-imp	81.61	57.26	67.30
	-soc	80.49	53.23	64.08
	-temp	74.51	61.29	67.26
	-spec	71.03	61.29	65.80
	-alt	75.53	57.26	65.14
	-con	79.57	59.68	68.20
	-goal	64.89	68.55	66.67

Level-of-Detail	*-pol*	46.62	59.62	52.32
	-bop	51.47	50.48	50.97
	-imp	49.79	57.21	53.24
	-soc	49.32	51.92	50.59
	-temp	49.79	55.77	52.61
	-spec	49.43	62.02	55.01
	-alt	47.39	56.73	51.64
	-con	55.61	54.81	55.21
	-goal	52.38	47.60	49.87
Manner	*-pol*	66.67	47.06	55.17
	-bop	69.23	52.94	60.00
	-imp	80.00	47.06	59.26
	-soc	100.00	47.06	64.00
	-temp	63.64	41.18	50.00
	-spec	80.00	47.06	59.26
	-alt	72.73	47.06	57.14
	-con	66.67	47.06	55.17
	-goal	68.75	64.71	66.67
Purpose	*-pol*	86.87	96.63	91.49
	-bop	91.40	95.51	93.41
	-imp	88.54	95.51	91.89
	-soc	90.53	96.63	93.48
	-temp	89.47	95.51	92.39
	-spec	92.47	96.63	94.51
	-alt	87.00	97.75	92.06
	-con	93.41	95.51	94.44
	-goal	92.39	95.51	93.92
Substitution	*-pol*	40.74	42.31	41.51
	-bop	50.00	50.00	50.00
	-imp	45.45	38.46	41.67
	-soc	48.00	46.15	47.06
	-temp	41.38	46.15	43.64
	-spec	47.83	42.31	44.90
	-alt	45.00	34.62	39.13
	-con	51.85	53.85	52.83
	-goal	50.00	53.85	51.85
Synchronous	*-pol*	46.67	20.00	28.00
	-bop	45.00	25.71	32.73
	-imp	44.44	22.86	30.19
	-soc	47.62	28.57	35.71
	-temp	22.22	11.43	15.09
	-spec	64.29	25.71	36.73
	-alt	30.77	22.86	26.23
	-con	43.75	20.00	27.45
	-goal	43.75	20.00	27.45

Table 27: Ablation studies for PDTB implicit DR classification, based on *InputForRelCls*. Similar to RST, lowest F1 scores are shown in blue, with the exception of *Cause+Belief*, for which removing the majority of UDims yields 00.00.