CVAE-based Re-anchoring for Implicit Discourse Relation Classification

Zujun Dou, Yu Hong*, Yu Sun, Guodong Zhou
School of Computer Science and Technology, Soochow University, China
{douzujun, tianxianer, sunyu41679@gmail}@gmail.com
gdzhou@suda.edu.cn

Abstract

Training implicit discourse relation classifiers suffers from data sparsity. Variational AutoEncoder (VAE) appears to be the proper solution. It is because that VAE is able to automatically generate inexhaustible varying samples by self supervision, and facilitates data augmentation. However, our experiments show that the utilization of VAE results in severe performance degradation. We ascribe this phenomenon to erroneous sampling. To address the issue, we use Conditional VAE (CVAE) to estimate the risk of erroneous sampling. Moreover, we develop a re-anchoring method which migrates the anchor of sampling area of VAE to reduce the risk. The experiments on PDTB v2.0 demonstrate that, compared to the RoBERTa-based baseline, re-anchoring yields substantial improvements. In addition, we prove that re-anchoring can cooperate with other auxiliary strategies (transfer learning and interactive attention mechanism) to further improve the classification performance.

1 Introduction

Implicit discourse relation classification is a task of determining relationships between arguments without connectives. We provide an example in Appendix A. Due to the omission of connectives (Zhou et al., 2010), classifying relations heavily relies on recognizable representations of arguments.

Learning richer and diverse linguistic phenomena from a large number of samples (relation-aware argument pairs) helps to enhance encoding, producing more recognizable representations (Ruan et al., 2020). However, there is lack of labeled data for learning. To overcome the bottleneck, previous studies have explored two classes of methods. Some of them conducted data expansion using PDTB explicit samples (Marcu and Echihabi, 2002; Braud and Denis, 2014; Rutherford and Xue, 2015; Xu et al., 2018) and parallel corpora (Wu et al., 2016; Shi et al., 2017, 2018). Others dug deeper into the existing data (instead of expanding it) to squeeze out additional salient features, where implicit connectives are speculated and annotated, and predicting them by machine is used as a supplementary task in a multi-task learning architecture (Qin et al., 2017; Shi and Demberg, 2019).

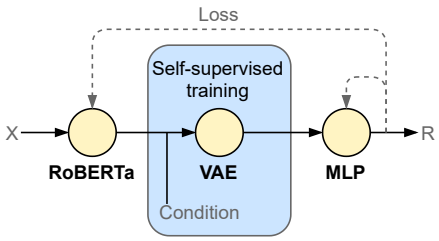

Figure 1: The three-stage encoder community.

In this paper, we attempt to enhance representation learning without using any external resources or artificially-created implicit connectives. We couple VAE (Kingma and Welling, 2014) with RoBERTa (Liu et al., 2019) and MLP to build a three-stage encoder community (Figure 1). It is inspired by the ability of VAE in generating variants (Section 2), and more importantly, the variants cover a wider range of linguistic phenomena. Specially, we utilize VAE to generate numerous variants for initial representations of arguments, and use them to challenge both RoBERTa and MLP (Section 3). Ideally, this helps to make the encoder community generalize well.

However, the use of VAE is proven ineffective in our experiments. It performs much worse than the less-sophisticated model that simply couples RoBERTa with MLP. Data analysis shows that the main drawback is caused by erroneous sampling. The errors occur when VAE tends to produce quite unusual variants (Section 4). To address the issue, we propose a re-anchoring strategy (Section 5) to migrate potential variants from high-risk sampling areas to low-risk. Instead of VAE, CVAE (Sohn

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1275–1283
November 7–11, 2021.

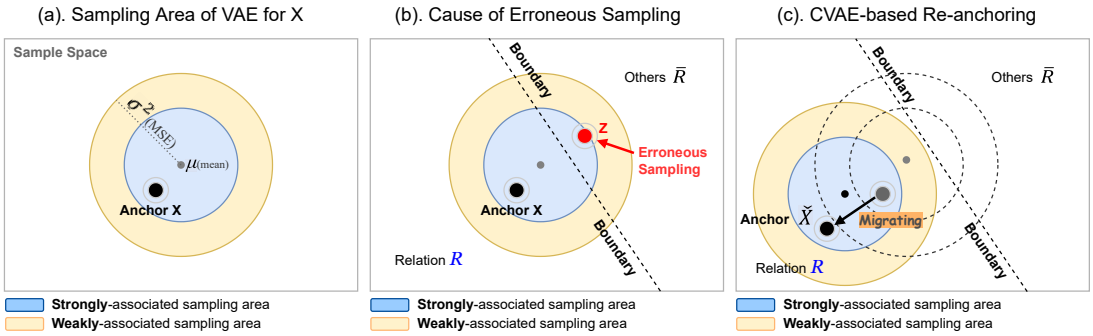

Figure 2: The schematic diagrams regarding sampling area, erroneous sampling and re-anchoring.

et al., 2015) is used for re-anchoring. We experiment on PDTB v2.0 (Prasad et al., 2008). Experimental results (Section 6) show that re-anchoring yields significant performance advantages. More importantly, it is proven that the cooperation between re-anchoring and other auxiliary strategies (transfer learning and interactive attention mechanism) yields further improvements.

2 Variational AutoEncoder (VAE)

VAE is an encoder which produces the hidden variable Z=$\{z_1, ...,z_n\}$ ($z_i \in \mathbb{R}$) for the input representation X=$\{x_1, ...,x_n\}$ ($x_i \in \mathbb{R}$). The variable is obtained by random sampling, from the finite sampling area where all samples are distributed with the posterior probability $p(Z|X)$.

During training, VAE plays an adversarial game with a decoder as below. VAE originally prefers to sample the variables Zs that are different from X. Therefore, it computes $p(Z|X)$ to approximate less-associated probability distributions with X. However, the decoder tends to completely reconstruct X using Zs. Therefore, it requests VAE to compromise, sampling similar Zs by approximating strongly-associated distributions with X. As a result, VAE learns to generate various different-but-similar representations Zs for X. All in all, VAE regards the input representation X as an anchor in the sample space, and estimates the sampling area around the anchor where, as shown in graph (a) in Figure 2, some samples are of strongly-associated distributions, others weakly-associated.

In our experiments we set $p(Z|X)$ to Gaussian distribution function $G(Z|X, \mu, \sigma^2)$, where μ denotes the geometric mean, while σ^2 the square deviation. In this case, the samples which are distributionally similar to μ will be more easily sampled and operationalized as Zs. In addition, we combine BiLSTM (Graves and Schmidhuber, 2005) with CNN (Zhang and Wallace, 2016) to build VAE, which serves to predict μ and σ^2 conditioned on the input X. The decoder is a BiLSTM unit.

3 Three-stage Encoder Community

As shown in Figure 1, we carry out a three-stage encoding process for arguments Arg 1 and Arg2. First, the arguments are fed into RoBERTa with the standard input format: [*CLS*]*Arg1*[*SEP*]*Arg2*[*EOS*]. The *CLS* embedding (*CLS* for short) output by RoBERTa is used as the initial representation. It contains the self-attentive information of both arguments. At the second stage, *CLS* is input into VAE. Using *CLS* as the anchor (X=*CLS*), VAE estimates the sampling area, and conducts random sampling in the area to produce the hidden variable Z. At the final stage, MLP is utilized to encode Z, computing the final representation $\hat{Z}$ of the arguments. Conditioned on $\hat{Z}$, the fully-connected layer with Softmax normalization predicts the implicit relation of the arguments.

Due to the addition of VAE in the middle, the MLP encoder that operates at the subsequent stage will encounter protean representations Zs of a single pair of arguments, during all the training epochs (one Z per epoch). Ideally, this should produce the effect of data augmentation, and thus strengthens the representation learning of the MLP encoder. However, the fact remains that VAE results in performance degradation during the test (Section 6).

4 Erroneous Sampling

The primary drawback of using VAE for data augmentation is erroneous sampling. It is caused by the following two reasons:

- The sampling area of VAE lies across the class boundary. As shown in graph (b) in Figure 2,

part of the samples in the area are actually heterogeneous with the anchor X (occurring at the other side of the boundary). In our study, it means that such samples hold different classes of relations from the anchor.

- By random sampling, the heterogeneous samples may be taken. Thus, they are mistakenly regarded as the family of the anchor, and used to challenge MLP with an incorrect class label during training (i.e., the one of the anchor X).

We suggest that erroneous sampling is most probably a common phenomenon in the study on PDTB v2.0. In Appendix B, we explain the cause of the phenomenon and provide a variety of examples.

5 Re-anchoring by Conditonal VAE

To relieve erroneous sampling, we develop a re-anchoring strategy to migrate the anchor away from the class boundary. Conditional VAE (CVAE) is utilized for re-anchoring. It uses the relation types as the subsidiary conditions to constrain the encoding process of VAE.

Assume $\mathcal{R}_t$ denotes the relation type that is held by the argument pair t. It is represented by the embedding $\mathcal{B}_t$, which is obtained by random initialization and element-wise accumulation with the unit vector. On the basis, given CLS_t that is output by RoBERTa for t, we combine $\mathcal{B}_t$ with CLS_t (by element-with accumulation) to form the input $\check{X}_t$ of VAE: $\check{X}_t$=$CLS_t \bigoplus \mathcal{B}_t$. This input appears as a new anchor migrating from the original position towards the relation-type embedding $\mathcal{B}_t$. Conditioned on this new anchor, the sampling area will be re-estimated by VAE in the region near $\mathcal{B}_t$ of R_t (See graph (c) in Figure 2). From the perspective of the spatial position in the entire sample space, the re-estimated sampling area is pulled away from the class boundaries, more or less. This reduces the risk of erroneous sampling.

Nevertheless, CVAE cannot be directly used during test because the relation type of every pair of arguments are unavailable at the moment (viz., it is an object needs to be predicted during test instead of being used as the prior knowledge). We made a detour, driving RoBERTa to learn re-anchoring.

Assume $\mathcal{R}_t$ and $\bar{\mathcal{R}}_t$ denotes the relation of the argument pair t and other relations, respectively. Both of their embeddings $\mathcal{B}_t$ and $\bar{\mathcal{B}}_t$ are obtained by random initialization. Though, to distinguish between them, $\mathcal{B}_t$ is combined with a unit vector.

MODEL	COM	CON	EXP	TEM
Baseline	53.71	59.30	75.90	32.46
Baseline+VAE	48.22	56.93	70.07	28.25
+Re-anchoring	**56.60**	62.60	77.74	37.13
+Transfer	54.85	59.52	79.63	40.16
+Attention	55.52	62.03	78.17	34.09
+ALL	55.72	**63.39**	**80.34**	**44.01**

Table 1: Results of ablation experiments (Binary classification is considered for each main relation class, and F1-score (%) is used as the evaluation metric).

On the basis, we feed $\check{X}_t$ ($\check{X}_t$=$CLS_t \bigoplus \mathcal{B}_t$) into VAE and use it to produce a variant $\mathcal{V}_t$ of t. Using $\mathcal{V}_t$, we estimate the risk $\mathcal{L}_\mathcal{B}$ of erroneous sampling:

$$\mathcal{L}_\mathcal{B} = \alpha f(\mathcal{V}_t, \mathcal{B}_t) - \beta f(\mathcal{V}_t, \bar{\mathcal{B}}_t) \quad (1)$$

where, f denotes the mean-square deviation function. It estimates the divergence between embeddings. Besides, α and β are hyperparameters. The risk $\mathcal{L}_\mathcal{B}$ will be enlarged when the sampled variant $\mathcal{V}_t$ is closer to $\bar{\mathcal{B}}_t$ but far from $\mathcal{B}_t$. In other word, once $\mathcal{V}_t$ has a small divergence with the embedding $\bar{\mathcal{B}}_t$ of other relations $\bar{\mathcal{R}}_t$, but large with that ($\mathcal{B}_t$) of the true relation $\mathcal{R}_t$, the risk $\mathcal{L}_\mathcal{B}$ will be high.

We introduce such a risk into the computation of the classification loss during the training of our three-stage encoder: Loss=$\mathcal{L}_\mathcal{C}$+$\mathcal{L}_\mathcal{B}$, where $\mathcal{L}_\mathcal{C}$ is the cross-entropy classification loss and $\mathcal{L}_\mathcal{B}$ the risk of erroneous sampling. Note that CVAE is a self-supervised model, and therefore it is independent of the training process of the classifier. Thus, the loss merely influences RoBERTa and MLP when the back propagation (BP) algorithm runs. Considering that MLP (at the final encoding stage) has nothing to do with the cause of the risk $\mathcal{L}_\mathcal{B}$, we suggest that propagating the risk has positive effects merely upon RoBERTa (the first-stage encoder) during BP. This facilitates RoBERTa to learn re-anchoring, so as to help CVAE estimate low-risk sampling area. During test, we couple the fine-tuned RoBERTa with VAE instead of CVAE. In this case, the relation type of every argument pair is masked.

6 Experimentation

6.1 Dataset, Evaluation and Settings

We experiment on the benchmark dataset PDTB v2.0 (Prasad et al., 2008). For comparison purpose, we select sections 02-20 in it as the training set, sections 00-01 the development, and sections 21-

Method	COM	CON	EXP	TEM	4-way F1	4-way *Acc.*
Zhang et al. (2015)	33.22	52.04	69.59	30.54		
Chen et al. (2016)	40.17	54.76		31.32		
Qin et al. (2016)	41.55	57.32	71.50	35.43		
Liu et al. (2016)	37.91	55.88	69.97	37.17	44.98	57.27
Liu and Li (2016)	36.70	54.48	70.43	38.84	46.29	57.17
Qin et al. (2017)	40.87	54.56	72.38	36.20		
Lan et al. (2017)	40.73	58.96	72.47	38.50	47.80	57.39
Bai and Zhao (2018)	47.85	54.47	70.60	36.97	51.06	
Guo et al. (2018)	40.35	56.81	72.11	38.65	47.59	59.06
Lei et al. (2018)	43.24	57.82	72.88	29.10	47.15	
Dai and Huang (2018)	46.79	57.09	70.41	45.61	48.82	57.44
Nguyen et al. (2019)	48.44	56.84	73.66	38.60	53.00	
Varia et al. (2019)	44.10	56.02	72.11	44.41	50.20	59.13
He et al. (2020)	47.98	55.62	69.37	38.94	51.24	59.94
Liu et al. (2020)	**59.44**	60.98	77.66	**50.26**	63.39	69.06
Ours	55.72	**63.39**	**80.34**	44.01	**65.06**	**70.17**

Table 2: Comparison to state-of-the-art methods. **Our** model is the one which strengthens the three-stage encoder community by re-anchoring, transfer learning and interactive attention. F1-score (%) and Acc (%) are used.

22 the test. The statistics of instances in them are presented in Appendix D. We use F1-score and Accuracy ($Acc.$) as the evaluation metrics. The settings of hyperparameters are detailed in Appendix D. Specially, both the risk trade-off factors α and β in equation (1) are set to 0.5.

6.2 Details of VAE (Input, Architecture, Computation and Training)

The input is formed by [CLS]Arg1[Sep]Arg2[EOS] output by RoBERTa. Both [CLS] and [Sep] serve as a 768-dimensional vector, which are the same with that of each token in the arguments Arg1 and Arg2. The length of each Argument is set to 126. Padding is used.

Our VAE comprises BiLSTM and CNN. BiLSTM predicts hidden states for every token in the input (including [CLS], [Sep] and all words in Arguments, one token per timestep). VAE outputs 256 768-dimensional vectors, which is used as a 256$*$768 matrix. Such a matrix is fed into CNN, a network comprising two groups of filters in the size of 2$*$768 and 4$*$ 768 respectively. Each group contains 128 filters. Using CNN along with 2 linear FC layers, we convolute the input matrix into a pair of 128$*$ 768 matrices (where, padding and dropout operations are used while pooling is not used), and concatenate them to form a 256$*$ 768 matrix. We split the matrix into two 256$*$384 submatrices. One of them is used to represent the independent variables μ of the Gaussian distributions, the other the σ. On the basis, re-parameterization is conducted when sampling. Re-parameterization contributes to the acquisition of non-negative variances.

When the encoder community is trained for relation classification, the relation embeddings are kept unchanged. RoBERTa is finetuned, though this is independent of self-supervised learning of VAE. VAE is trained separately, with the goal of reconstructing the input well. We shut down the training course when the observable development performance is going to be steady. During test taking, the relation embeddings are disable.

6.3 Results and Analysis

In the ablation experiments, we examine the binary classification performance for the four main relation types, including `Comparison` (COM), `Contingency` (CON), `Expansion` (EXP) and `Temporality` (TEM). The joint model that assembles RoBERTa and MLP is taken as the baseline. We improve the baseline by the following auxiliary strategies: 1) coupling it with **VAE** (i.e., forming the three-stage encoder community); 2) conducting **re-anchoring** when applying VAE during test; 3) additionally equipping VAE with interactive **attention** mechanism (Ruan et al., 2020); 4) retraining RoBERTa by **transfer** learning (Nie et al., 2019) upon the explicit discourse relation dataset (Prasad et al., 2008), and 5) employing **all** the above strategies to form a cooperation model.

We show the test results in Table 1. It can be observed that simply coupling VAE with the baseline actually results in a severe performance degradation for all the considered relation classes, yielding much lower F1-scores than the baseline. By

contrast, re-anchoring substantially improves the baseline. More importantly, it has a comparable performance to transfer learning, although the latter additionally uses a considerable amount of external data (18,459 explicitly-related argument pairs). In addition, the cooperation of all the auxiliary strategies achieves the best performance. It is noteworthy that Ruan et al. (2020)'s interactive attention is used in the cooperation model, and it is necessarily equipped with VAE from behind. This is because the random sampling stage of VAE invalidates the self-attention mechanism of RoBERTa. The supplementary interactive attention mechanism helps to recover the distracted attention. As shown in Table 1, it yields a considerable improvement.

We compare the cooperation model to the state of the art. As shown in Table 2, it achieves the best performance (F1-score and Acc) for the 4-way classification of all the considered relation classes. Moreover, it has a comparable performance to Liu et al. (2020)'s RoBERTa-based context-aware multi-perspective fusion model, in the binary classification scenarios (one relation class vs others).

6.4 Case Study

We verify the effectiveness of re-anchoring by measuring the percentage of salvaging the mistakenly-determined semantically-similar argument pairs.

Given two groups of argument pairs (i.e., two pairs of arguments, which comprise 4 arguments in total), we present each of them (i.e., one argument pair) using the [CLS] embeddings of the arguments. Concatenation is used. On the basis, we calculate the Cosine similarity between the [CLS]-based representations of the two groups of argument pairs. A empirically-set threshold is adopted as the condition during the time when we determine the similarity. The two groups of argument pairs is determined as semantically similar when the cosine similarity is larger than the threshold, otherwise dissimilar.

There are 5,165 groups of semantically-similar argument pairs found in the test set, each of which hold different types of relations. Within them, there are 2,685 cases were incorrectly determined for binary relation classification by the VAE-based baseline, occupying 52% of the total examples, while 2,220 for 4-way classification, occupying 43%. By contrast, the CVAE-based re-anchoring salvaged 1,962 and 1,601 cases for binary and 4-way classification respectively, occupying 38% and 31% of the mistakenly determined cases.

7 Conclusion

We develop a three-stage encoder community for implicit relation recognition. VAE is used for data augmentation. In particular, we propose a CVAE-based re-anchoring strategy to solve the problem of erroneous sampling. Experimental results show that our method yields substantial improvements.

Data analysis demonstrates that a PDTB argument may be accompanied with two different arguments, bringing inconsistent relations. In the future, we will develop a context-aware adversarial model to selectively assign attention weights to the central argument, conditioned on both companions.

8 Acknowledgement

We thank all reviewers for their insightful comments, as well as the great efforts our colleagues have made so far. This work is supported by the national Natural Science Foundation of China (NSFC) and Major National Science and Technology project of China, via Grant Nos. 62076174 and 2020YBF1313601.

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *COLING*, pages 571–583.

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1694–1705, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *ACL*, pages 1726–1735.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *NAACL-HLT*, pages 141–151.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *COLING*, pages 547–558.

Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han. 2020. Transs-driven joint learning architecture for implicit discourse relation recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 139–148.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *EMNLP*, pages 1299–1308.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *AAAI*, pages 4848–4855.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3830–3836. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *EMNLP*, pages 1224–1233.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *AAAI*, pages 2750–2756.

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.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 368–375.

Linh The Nguyen, Ngo Van Linh, Khoat Than, and Thien Huu Nguyen. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *ACL*, pages 4201–4207.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *EMNLP*, pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1006–1017. Association for Computational Linguistics.

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

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.

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

Wei Shi, Frances Yung, and Vera Demberg. 2018. Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. *CoRR*, abs/1808.10290.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction:

Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden. Association for Computational Linguistics.

Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. 2016. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2306–2312, Austin, Texas. Association for Computational Linguistics.

Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 725–731, Brussels, Belgium. Association for Computational Linguistics.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *EMNLP*, pages 2230–2235.

Ye Zhang and Byron Wallace. 2016. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *COLING*, pages 1507–1514.

Appendix.

A Definition and Example of Implicit Discourse Relation

Implicit discourse relation classification is a task of determining relationships between arguments, where the connective of the arguments fails to be explicitly given. Each argument is either a sentence or clause. The connective refers to a conjunction which explicitly signals the relation.

For example, the two arguments in (1) hold an implicit `Causal` relation instead of explicit because the possible connective "*because*" is omitted.

(1) **Arg1**: *Psyllium's not a good crop.*
Arg2: *You get a rain at the wrong time and the crop is ruined.*

where, `Causality` stands for a subtype relation of the major relation `Contingency`. In our experiments, the four major relation types are considered, including `Expansion`, `Contingency`, `Comparison` and `Temporality`.

B High Occurrence Rate of Erroneous Sampling in PDTB

Erroneous sampling occurs frequently when VAE performs on PDTB v2.0. It is because that a large number of pairwise arguments in the corpus are selected from the same discourses. Unavoidably, some of them are similar in the use of words or present similar semantics, though they hold different types of relations. As a result, the CLS embeddings (anchors) of them (derived from RoBERTa) are distributed near the class boundary. Therefore, the proportion of the sampling area that spreads across the class boundary is considerably large. Even, the strongly-associated distribution area may spread across the boundary. This aggravates erroneous sampling.

For example, the two pairs of arguments in (2) and (3) are taken from the same PDTB document (ID: wsj_ 0045). They are constituted with a number of the same words. More importantly, they are of similar semantics, to some extent. However, the types of the relations they hold are different. One of them is `Contingency`, the other `Comparison`.

(2) **Arg1**: *When Scoring High first came out in 1979, it was a publication of Random House.*
Arg2: *McGraw-Hill was outraged.*
[Relation: **Contingency**]

Relation Type	Training	Dev	Test
Comparison (**COM**)	1,855	189	145
Contingency (**CON**)	3,235	281	273
Expansion (**EXP**)	6,673	638	538
Temporiality (**TEM**)	582	48	55
Total	12,345	1,156	1,011

Table 3: Data statistics in PDTB sets.

(3) **Arg1**: *But in 1988, McGraw-Hill purchased the Random House unit that publishes Scoring High.*
Arg2: *they are unaware of any efforts by McGraw-Hill to modify or discontinue Scoring High.*
[Relation: **Comparison**]

There are additional 6 groups of examples exhibited in Table 4, where the semantic-level similarity between argument pairs is given. The similarity is calculated by Cosine function upon the CLS embeddings of argument pairs. The CLS embeddings are obtained using the pre-trained language model ReBERTa. We collect all the cases from PDTB which has a similarity higher than 9.8 and attach them with this submission.

C Statistics in PDTB

We experiment on PDTB v2.0 (Prasad et al., 2008). The corpus comprises the ground-truth annotations of implicit discourse relations for 12,345 argument pairs. The argument pairs are assigned to 23 sections. For comparison purpose, we follow the common practice to use the dataset, selecting sections 02-20 as the training set, sections 00-01 the development (Dev), and sections 21-22 the test. Table 3 shows the statistics in the sets, as well as the sample distributions of the four main relation classes.

D Hyperparameter Settings

The dimension of each hidden state output by RoBERTa is set to 768 (d=768). The length of arguments is uniformly set to 126 ($n1$=$n2$=126), and the length of the combined argument representation is set to 256 (252 plus 4 separators). Each of the BiLSTM units in VAE is of 256 dimensions (d_h=256). Finally, we set the filter region size of convolution kernel of CNN units in VAE as (2, 4).

During training, we set the mini-batch size to 8 (argument pairs) and specify the dropout rate as 0.2. We set the learning rate to 5e-6.

Doc-ID	Score	Argument Pairs	Relation
wsj_0003	0.99922	**Arg1** About 160 workers at a factory that made paper for the Kent filters were exposed to asbestos in the 1950s, **Arg2** Areas of the factory were particularly dusty where the crocidolite was used.	**Expansion**
		Arg1 Areas of the factory were particularly dusty where the crocidolite was used, **Arg2** Workers dumped large burlap sacks of the imported material into a huge bin, poured in cotton and acetate fibers and mechanically mixed the dry fibers in a process used to make filters.	**Contingency**
wsj_0010	0.99889	**Arg1** The next morning, with a police escort, busloads of executives and their wives raced to the Indianapolis Motor Speedway, **Arg2** so the lieutenant governor welcomed the special guests.	**Temporality**
		Arg1 After the race, Fortune 500 executives drooled like schoolboys over the cars and drivers, **Arg2** No dummies, the drivers pointed out they still had space on their machines for another sponsor's name or two.	**Contingency**
wsj_0018	0.99937	**Arg1** Cray Research's decision to link its $98.3 million promissory note to Mr. Cray's presence will complicate a valuation of the new company, **Arg2** It has to be considered as an additional risk for the investor.	**Expansion**
		Arg1 It has to be considered as an additional risk for the investor, **Arg2** Cray Computer will be a concept stock.	**Contingency**
wsj_0051	0.99872	**Arg1** The Ministry of International Trade and Industry summoned executives from the companies to "make sure they understood" the concern about such practices, according to a government spokesman, **Arg2** These cases lead to the loss of the firms' social and international credibility.	**Contingency**
		Arg1 The fire is also fueled by growing international interest in Japanese behavior, **Arg2** So far there have been no public overseas complaints about the issue.	**Comparison**
wsj_0059	0.99740	**Arg1** Dollar-yen trade is the driving force in the market but I'm not convinced it will continue, **Arg2** Who knows what will happen down the road, in three to six months, if foreign investment starts to erode.	**Contingency**
		Arg1 In late New York trading yesterday, the dollar was quoted at 1.8500 marks, up from 1.8415 marks late Tuesday, and at 143.80 yen, up from 142.85 yen late Tuesday, **Arg2** Sterling was quoted at $1.5755, down from $1.5805 late Tuesday.	**Expansion**
wsj_0063	0.99857	**Arg1** In May, the two companies, through their jointly owned holding company, Temple, offered $50 a share, or $777 million, **Arg2** In August, Temple sweetened the offer to $63 a share, or $963 million.	**Temporality**
		Arg1 That $130 million gives us some flexibility in case Temple raises its bid, **Arg2** We are able to increase our price above the $70 level if necessary.	**Expansion**

Table 4: Examples regarding the cause of erroneous sampling. The pairs of arguments which lie in the same discourse (document) may be semantically similar though holding different types of relations.(Note: The column of Doc-ID shows the document number, while score denotes the similarity score calculated by Cosine function over CLS embeddings.)