Distributed Marker Representation for Ambiguous Discourse Markers and Entangled Relations

Dongyu Ru[1], **Lin Qiu**[1], **Xipeng Qiu**[2], **Yue Zhang**[3], **Zheng Zhang**[1]
[1]Amazon AWS AI [2]School of Computer Science, Fudan University
[3]School of Engineering, Westlake University
{rudongyu,quln,zhaz}@amazon.com
xpqiu@fudan.edu.cn
zhangyue@westlake.edu.cn

Abstract

Discourse analysis is an important task because it models intrinsic semantic structures between sentences in a document. Discourse markers are natural representations of discourse in our daily language. One challenge is that the markers as well as pre-defined and human-labeled discourse relations can be ambiguous when describing the semantics between sentences. We believe that a better approach is to use a contextual-dependent distribution over the markers to express discourse information. In this work, we propose to learn a Distributed Marker Representation (DMR) by utilizing the (potentially) unlimited discourse marker data with a latent discourse sense, thereby bridging markers with sentence pairs. Such representations can be learned automatically from data without supervision, and in turn provide insights into the data itself. Experiments show the SOTA performance of our DMR on the implicit discourse relation recognition task and strong interpretability. Our method also offers a valuable tool to understand complex ambiguity and entanglement among discourse markers and manually defined discourse relations.

1 Introduction

Discourse analysis is a fundamental problem in natural language processing. It studies the linguistic structures beyond the sentence boundary and is a component of chains of thinking. Such structural information has been widely applied in many downstream applications, including information extraction (Peng et al., 2017), long documents summarization (Cohan et al., 2018), document-level machine translation (Chen et al., 2020), conversational machine reading (Gao et al., 2020).

Discourse relation recognition (DRR) focuses on semantic relations, namely, discourse senses between sentences or clauses. Such inter-sentence structures are sometimes explicitly expressed in natural language by discourse connectives, or markers (e.g., *and*, *but*, *or*). The availability of these markers makes it easier to identify corresponding relations (Pitler et al., 2008), as is in the task of explicit discourse relation recognition (EDRR), since there is strong correlation between discourse markers and relations. On the contrary, implicit discourse relation recognition (IDRR), where markers are missing, remains a more challenging problem.

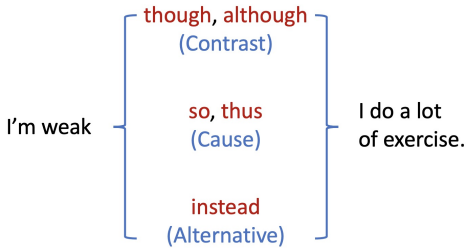


Figure 1: Entangled discourse relations and corresponding markers between clauses. As shown in the figure, there exist diverse discourse relations (marked in blue) and corresponding markers (marked in red) for the same pair of clauses. It suggests that the semantic meaning of different discourse relations can be entangled to each other.

Prior work aims to address such challenges by making use of discourse marker information over explicit data in learning implicit discourse relations, either by injecting marker prediction knowledge into a representation model (Zhou et al., 2010; Braud and Denis, 2016), or transferring the marker prediction task into implicit discourse relation prediction by manually defining a marker-relation mapping (Xiang et al., 2022; Zhou et al., 2022). It has been shown that discourse marker information can effectively improve relation prediction results. Nevertheless, relatively little work has investigated various subtleties concerning the correlation between discourse markers and discourse relations, and their effect to IDRR in further detail.

To properly model discourse relations and markers, we need to consider that manually-defined discourse relations can be semantically entangled and

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5334–5351

markers are ambiguous. As shown in Fig. 1, for a pair of clauses, based on different emphasis on semantics, we have different choices on discourse relations and their corresponding markers. The existence of multiple plausible discourse relations indicates the entanglement between their semantic meaning. Besides, discourse markers and relations do not exclusively map to each other. As an example, "Ann went to the movies, **and** Bill went home" (*Temporal.Synchrony*) and "Ann went to the movies, **and** Bill got upset" (*Contingency.Cause*) both use the marker **and** but express different meanings. Identifying relations based on single markers are difficult in certain scenarios because of such ambiguity. Thus, a discrete and deterministic mapping between discourse relations and markers can not precisely express the correlations between them.

Based on the study of above issues, we propose to use **D**istributed **M**arker **R**epresentation to enhance the informativeness of discourse expression. Specifically, We use a probabilistic distribution on markers or corresponding latent senses instead of a single marker or relation to express discourse semantics. We introduce a bottleneck in the latent space, namely a discrete latent variable indicating discourse senses, to capture semantics between clauses. The latent sense then produces a *distribution* of plausible markers to reflect its surface form. This probabilistic model, which we call DMR, naturally deals with ambiguities between markers and entanglement among the relations. We show that the latent space reveals a hierarchical marker-sense clustering, and that entanglement among relations are currently under-reported. Empirical results on the IDRR benchmark Penn Discourse Tree Bank 2 (PDTB2) (Prasad et al., 2008) shows the effectiveness of our framework. We summarize our contributions as follows:

- We propose a latent-space learning framework for discourse relations and effectively optimize it with cheap marker data.[1]

- With the latent bottleneck and corresponding probabilistic modeling, our framework achieves the SOTA performance on implicit discourse relation recognition without a complicated architecture design.

- We investigate the ambiguity of discourse markers and entanglement among discourse relations to explain the plausibility of probabilistic modeling of discourse relations and markers.

[1]Code is publicly available at: https://github.com/rudongyu/DistMarker

2 Related Work

Discourse analysis (Brown et al., 1983; Joty et al., 2019; McCarthy et al., 2019), targets the discourse relation between adjacent sentences. It has attracted attention beyond intra-sentence semantics. It is formulated into two main tasks: explicit discourse relation recognition and implicit discourse relation recognition, referring to the relation identification between a pair of sentences with markers explicitly included or not. While EDRR has achieved satisfactory performance (Pitler et al., 2008) with wide applications, IDRR remains to be challenging (Pitler et al., 2009; Zhang et al., 2015; Rutherford et al., 2017; Shi and Demberg, 2019). Our work builds upon the correlation between the two critical elements in discourse analysis: discourse relations and markers.

Discourse markers have been used for not only marker prediction training (Malmi et al., 2018), but also for improving the performance of IDRR (Marcu and Echihabi, 2002; Rutherford and Xue, 2015) and representation learning (Jernite et al., 2017). Prior efforts on exploring markers have found that training with discourse markers can alleviate the difficulty on IDRC (Sporleder and Lascarides, 2008; Zhou et al., 2010; Braud and Denis, 2016). Compared to their work, we focus on a unified representation using distributed markers instead of relying on transferring from explicit markers to implicit relations. Jernite et al. (2017) first extended the usage of markers to sentence representation learning, followed by Nie et al. (2019); Sileo et al. (2019) which introduced principled pretraining frameworks and large-scale marker data. Xiang et al. (2022); Zhou et al. (2022) explored the possibility of connecting markers and relations with prompts. In this work, we continue the line of improving the expression of discourse information as distributed markers.

3 Distributed Marker Representation Learning

We elaborate on the probabilistic model in Sec. 3.1 and its implementation with neural networks in Sec. 3.2. We then describe the way we optimize the model (Sec. 3.3).

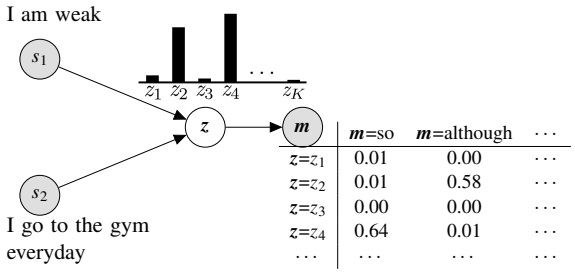


Figure 2: The graphical model of $p(\boldsymbol{m}|s_1, s_2)$. z is the latent variable indicating the latent sense, namely the semantic relation between two clauses. K is the number of candidate values for the random variable z.

3.1 Probabilistic Formulation

We learn the distributed marker representation by predicting markers given pairs of sentences. We model the distribution of markers by introducing an extra latent variable z which indicates the latent senses between two sentences. We assume the distribution of markers depends only on the latent senses, and is independent of the original sentence pairs when z is given, namely $\boldsymbol{m} \perp (s_1, s_2)|z$.

$$p_{\psi,\phi}(\boldsymbol{m}|s_1, s_2) = \sum_z p_\psi(z|s_1, s_2) \cdot p_\phi(\boldsymbol{m}|z) \quad (1)$$

where the latent semantic senses z describes the unambiguous semantic meaning of m in the specific context, and our target is to model the probabilistic distribution $p(\boldsymbol{m}|s_1, s_2)$ with z. The probabilistic model is depicted in Fig. 2 with an example.

The key inductive bias here is that we assume the distribution of discourse markers is independent of the original sentence pairs given the latent semantic senses (Eq. 1). This formulation is based on the intuition that humans decide the relationship between two sentences in their cognitive worlds first, then pick one proper expression with a mapping from latent senses to expressions (which we call *z2m* mapping in this paper) without reconsidering the semantic of sentences. Decoupling the *z2m* mapping from the distribution of discourse marker prediction makes the model exhibit more interpretability and transparency.

Therefore, the probabilistic distribution of $p_{\psi,\phi}(\boldsymbol{m}|s_1, s_2)$ can be decomposed into $p_\psi(m|z)$ and $p_\phi(z|s_1, s_2)$ based on the independence assumption above. ψ and ϕ denote parameters for each part [2]. The training objective with latent senses included is to maximize the likelihood on large-scale corpus under this assumption:

$$\mathcal{L}(\psi, \phi) = \mathbb{E}_{(s_1,s_2,m)\sim D} \log p_{\psi,\phi}(m|s_1, s_2). \quad (2)$$

3.2 Neural Architecture

Our model begins by processing each sentence with an encoder `SentEnc`:

$$h = \texttt{SentEnc}_{\psi_s}([s_1, \texttt{[SEP]}, s_2]), \quad (3)$$

where $h \in \mathbb{R}^d$ denote the sentence pair representation in d dimensions for s_1 and s_2. ψ_s are parameters of the sentence encoder. The encoder is instantiated as a pre-trained language model in practice.

Then we use two linear layers to map the pair representation h to the distribution of z as below:

$$h_z = \psi_{w1} \cdot h + \psi_{b1}, \quad (4)$$

$$p_\theta(z|s_1, s_2) = \texttt{softmax}(\psi_{w2} \cdot h_z + \psi_{b2}), \quad (5)$$

where $\psi_{w1} \in \mathbb{R}^{d\times 4d}, \psi_{b1} \in \mathbb{R}^d, \psi_{w2} \in \mathbb{R}^{K\times d}, \psi_{b2} \in \mathbb{R}^K$ are trainable parameters. K is the dimension of latent discourse senses.

The parameter ψ_{w2} not only acts as the mapping from representation h_z to z's distribution, but can also be seen as an embedding lookup table for the K values of z. Each row in ψ_{w2} is a representation vector for the corresponding value, as an anchor in the companion continuous space of z.

To parameterize the *z2m* mapping, the parameter $\phi \in \mathbb{R}^{K\times N}$ is defined as a probabilistic transition matrix from latent semantic senses z to markers m (in log space), where N is the number of candidate markers:

$$\log p_\phi(m|z) = \log \texttt{softmax}(\phi), \quad (6)$$

where $\psi = (\psi_s, \psi_{w1}, \psi_{b1}, \psi_{w2}, \psi_{b2}), \phi$ are the learnable parameters for parameterize the distribution $p_{\psi,\phi}(\boldsymbol{m}|s_1, s_2)$.

3.3 Optimization

We optimize the parameters ψ and ϕ with the classic EM algorithm due to the existence of the latent variable z. The latent variable z serves as a regularizer during model training. In the E-step of each iteration, we obtain the posterior distribution $p(z|s_1, s_2, m)$ according to the parameters in the current iteration $\psi^{(t)}, \phi^{(t)}$ as shown in Eq. 7.

[2] We omit the subscript of parameters ψ and ϕ in some expressions later for conciseness.

Algorithm 1 EM Optimization for Discourse Marker Training with Latent Senses

Initialize model parameters as ψ^0, ϕ^0.
while not converge **do** ▷ t-th iteration
  Sample a batch of examples for EM optimization.
  for each example (s_1, s_2, m) in the EM batch **do**
    Calculate and save the posterior $p(z|s_1, s_2, m)$ according to $\psi^{(t)}, \phi^{(t)}$.
  end for
  for each example (s_1, s_2, m) in the EM batch **do**
    Estimate $\mathbb{E}_{p(z|s_1,s_2,m)}[\log p_{\psi,\phi}(m, z|s_1, s_2)]$ according to $\psi^{(t)}, \phi^{(t)}$. ▷ E-step
  end for
  Update parameters ψ to $\psi^{(t+1)}$ in mini-batch with the gradient calculated as $\nabla_\psi \mathcal{L}(\psi, \phi^{(t)})$.
  Update parameters ϕ to $\phi^{(t+1)}$ according to the updated $\psi^{(t+1)}$ and the gradient $\nabla_\phi \mathcal{L}(\psi^{(t+1)}, \phi)$. ▷ M-step
end while

Based on our assumption that $\boldsymbol{m} \perp (s_1, s_2)|z$, we can get the posterior distribution:

$$\begin{aligned} p(z|s_1, s_2, m) &= \frac{p(m|s_1, s_2, z) \cdot p(z|s_1, s_2)}{p(m|s_1, s_2)} \\ &= \frac{p(m|z) \cdot p(z|s_1, s_2)}{p(m|s_1, s_2)} \\ &\propto p_{\psi^{(t)}}(z|s_1, s_2) \cdot p_{\phi^{(t)}}(m|z). \end{aligned} \quad (7)$$

In M-step, we optimize the parameters ψ, ϕ by maximizing the expectation of joint log likelihood on estimated posterior $p(z|s_1, s_2, m)$. The updated parameters $\psi^{(t+1)}, \phi^{(t+1)}$ for the next iteration can be obtained as in Eq. 8.

$$\psi^{(t+1)}, \phi^{(t+1)} = \arg\max_{\psi,\phi} \mathbb{E}_{p(z|s_1,s_2,m)}[\log p_{\psi,\phi}(m, z|s_1, s_2)]. \quad (8)$$

In practice, the alternative EM optimization can be costly and unstable due to the expensive expectation computation and the subtlety on hyperparameters when optimizing ψ and ϕ jointly. We alleviate the training difficulty by empirically estimating the expectation on mini-batch and separate the optimization of ψ and ϕ. We formulate the loss functions as below, for separate gradient descent optimization of ψ and ϕ:

$$\mathcal{L}(\psi, \phi^{(t)}) = \mathrm{KLDiv}(p(z|s_1, s_2, m), p_{\psi,\phi^{(t)}}(m, z|s_1, s_2)),$$
$$\mathcal{L}(\psi^{(t+1)}, \phi) = -\log p_{\psi^{(t+1)},\phi}(m|s_1, s_2),$$

where $\phi^{(t)}$ means the value of ϕ before the t-th iteration and $\psi^{(t+1)}$ means the value of ψ after the t-th iteration of optimization. KLDiv denotes the Kullback-Leibler divergence. The overall optimization algorithm is summarized in Algorithm 1.

4 Experiments

DMR adopts a latent bottleneck for the space of latent discourse senses. We first prove the effectiveness of the latent variable and compare against current SOTA solutions on the IDRR task. We then examine what the latent bottleneck learned during training and how it addresses the ambiguity and entanglement of discourse markers and relations.

4.1 Dataset

We use two datasets for learning our DMR model and evaluating its strength on downstream implicit discourse relation recognition, respectively. See Appendix A for statistics of the datasets.

Discovery Dataset (Sileo et al., 2019) is a large-scale discourse marker dataset extracted from commoncrawl web data, the Depcc corpus (Panchenko et al., 2018). It contains 1.74 million sentence pairs with a total of 174 types of explicit discourse markers between them. Markers are automatically extracted based on part-of-speech tagging. We use top-k accuracy **ACC@k** to evaluate the marker prediction performance on this dataaset. Note that we use explicit markers to train DMR but evaluate it on IDRR thanks to different degrees of verbosity when using markers in everyday language.

Penn Discourse Tree Bank 2.0 (PDTB2) (Prasad et al., 2008) is a popular discourse analysis benchmark with manually-annotated discourse relations and markers on Wall Street Journal articles. We perform the evaluation on its implicit part with 11 major second-level relations included. We follow (Ji and Eisenstein, 2015) for data split, which is widely used in recent studies for IDRR. **Macro-**

Model	Backbone	macro-F_1	ACC
IDRR-C&E (Dai and Huang, 2019)	ELMo	33.41	48.23
MTL-MLoss (Nguyen et al., 2019)	ELMo	-	49.95
BERT-FT (Kishimoto et al., 2020)	BERT	-	54.32
HierMTN-CRF (Wu et al., 2020)	BERT	33.91	52.34
BMGF-RoBERTa (Liu et al., 2021)	RoBERTa	-	58.13
MTL-MLoss-RoBERTa† (Nguyen et al., 2019)	RoBERTa	38.10	57.72
HierMTN-CRF-RoBERTa† (Wu et al., 2020)	RoBERTa	38.28	58.61
LDSGM (Wu et al., 2022)	RoBERTa	40.49	60.33
PCP-base (Zhou et al., 2022)	RoBERTa	41.55	60.54
PCP-large (Zhou et al., 2022)	RoBERTa	**44.04**	61.41
DMR-base$_{\text{w/o } z}$	RoBERTa	37.24	59.89
DMR-large$_{\text{w/o } z}$	RoBERTa	41.59	62.35
DMR-base	RoBERTa	42.41	61.35
DMR-large	RoBERTa	43.78	**64.12**

Table 1: Experimental Results of Implicit Discourse Relation Classification on PDTB2. Results with † are from Wu et al. (2022). DMR-large and DMR-base adopt roberta-large and roberta-base as `SentEnc`, respectively.

$\mathbf{F}_1$ and **ACC** are metrics for IDRR performance. We note that although annotators are allowed to annotate multiple senses (relations), only 2.3% of the data have more than one relation. Therefore whether DMR can capture more entanglement among relations is of interest as well (Sec. 4.5).

4.2 Baselines

We compare our DMR model with competitive baseline approaches to validate the effectiveness of DMR. For the IDRR task, we compare DMR-based classifier with current SOTA methods, including BMGF (Liu et al., 2021), which combines representation, matching, and fusion; LDSGM (Wu et al., 2022), which considers the hierarchical dependency among labels; the prompt-based connective prediction method, PCP (Zhou et al., 2022) and so on. For further analysis on DMR, we also include a vanilla sentence encoder without the latent bottleneck as an extra baseline, denoted as BASE.

4.3 Implementation Details

Our DMR model is trained on 1.57 million examples with 174 types of markers in Discovery dataset. We use pretrained RoBERTa model (Liu et al., 2019) as `SentEnc` in DMR. We set the default latent dimension K to 30. More details regarding the implementation of DMR can be found in Appendix A.

For the IDRR task, we strip the marker generation part from the DMR model and use the hidden state h_z as the pair representation. BASE uses the `[CLS]` token representation as the representation of input pairs. A linear classification layer is stacked on top of models to predict relations.

	BMGF	LDSGM	DMR
Comp.Concession	0.	0.	0.
Comp.Contrast	59.75	**63.52**	63.16
Cont.Cause	59.60	**64.36**	62.65
Cont.Pragmatic Cause	0.	0.	0.
Expa.Alternative	60.0	**63.46**	55.17
Expa.Conjunction	**60.17**	57.91	58.54
Expa.Instantiation	67.96	**72.60**	72.16
Expa.List	0.	8.98	**36.36**
Expa.Restatement	53.83	58.06	**59.19**
Temp.Async	56.18	56.47	**59.26**
Temp.Sync	0.	0.	0.
Macro-f1	37.95	40.49	**42.41**

Table 2: Experimental Results of Implicit Discourse Relation Recognition on PDTB2 Second-level Senses

4.4 Implicit Discourse Relation Recognition

We first validate the effectiveness of modeling latent senses on the challenging IDRR task.

Main Results DMR demonstrates comparable performance with current SOTAs on IDRR, but with a simpler architecture. As shown in Table 1, DMR leads in terms of accuracy by 2.7pt and is a close second in macro-F_1.

The results exhibit the strength of DMR by more straightforwardly modeling the correlation between discourse markers and relations. Despite the absence of supervision on discourse relations during DMR learning, the semantics of latent senses distilled by EM optimization successfully transferred to manually-defined relations in IDRR.

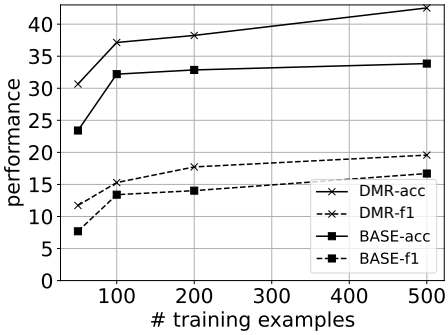


Figure 3: Few-shot IDRR Results on PDTB2

# Training Examples		25	100	500	full (10K)
BASE	ACC	-	32.20	33.85	59.45
	F1	-	13.40	16.70	34.34
$BASE_p$	ACC	-	33.76	37.56	60.90
	F1	-	13.54	17.21	35.45
$BASE_g$	ACC	19.12	34.07	39.23	**63.19**
	F1	5.75	13.72	19.27	36.59
DMR	ACC	**21.32**	**37.14**	**42.53**	62.97
	F1	**7.01**	**15.29**	**19.57**	**39.33**

Table 3: Few-shot IDRR Results on PDTB2

Model	ACC@1	ACC@3	ACC@5	ACC@10
Discovery	**24.26**	**40.94**	49.56	61.81
DMR_{30}	8.49	22.76	33.54	48.11
DMR_{174}	22.43	40.92	**50.18**	**63.21**

Table 4: Experimental results of marker prediction on the Discovery test set. DMR_{30} and DMR_{174} indicate the models with the dimension K equals to 30 and 174 respectively.

Marker	1st Cluster		2nd Cluster	
additionally	z_1:	as a result, in turn, simultaneously	z_{20}:	for example, for instance, specifically
amazingly	z_9:	thankfully, fortunately, luckily	z_{21}:	oddly, strangely, unfortunately
but	z_{19}:	indeed, nonetheless, nevertheless	z_{24}:	anyway, and, well

Table 5: Top 2 clusters of three random sampled markers. Each cluster corresponds to a latent z coupled with its top 3 markers.

Based on the comparison to DMR without latent z, we observe a significant performance drop resulted from the missing latent bottleneck. It indicates that the latent bottleneck in DMR serves as a regularizer to avoid overfitting on similar markers.

Fine-grained Performance We list the fine-granined performance of DMR and compare it with SOTA approaches on second-level senses of PDTB2. As shown in Table 2, DMR achieves significant improvements on relations with little supervision, like *Expa.List* and *Temp.Async*. The performance of majority classes, e.g. *Expa.Conjunction*, are slightly worse. It may be caused by the entanglement between *Expa.Conjunction* and *Expa.List* to be discussed in Sec. 4.5. In summary, DMR achieves better overall performance by maintaining equilibrium among entangled relations with different strength of supervision.

Few-shot Analysis Fig. 3 shows DMR achieves significant gains against BASE in few-shot learning experiments. The results are averaged on 3 independent runs for each setting. In fact, with only ~60% of annotated data, DMR achieves the same performance as BASE with full data by utilizing the cheap marker data more effectively.

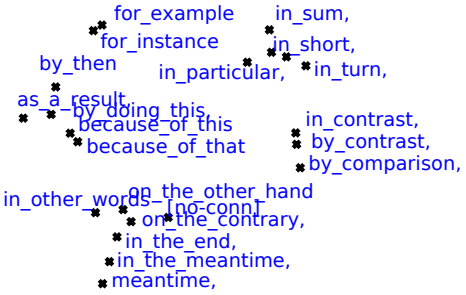
To understand the ceiling of the family of such BERT-based pretrained model with markers as an extra input, we augment the data in two ways: $BASE_g$ inserts the groundtruth marker, and $BASE_p$ where the markers are predicted by a model[3] officially released by Discovery (Sileo et al., 2019). Table 3 presents the results where the informative markers are inserted to improve the performance of BASE, following the observations and ideas from (Zhou et al., 2010; Pitler et al., 2008). DMR continues to enjoy the lead, even when the markers are groundtruth (i.e. $BASE_g$), suggesting DMR's hidden state contains more information than single markers.

4.5 Analysis & Discussion

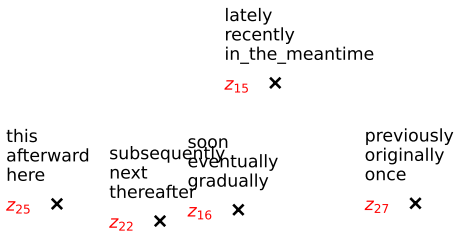
Marker Prediction The performance of DMR on marker prediction is sensitive to the capacity of the bottleneck. When setting K to be the number of markers (174), it matches and even outperforms the Discovery model which directly predicts the markers on the same data (Table 4). A smaller K sacrifices marker prediction performance but it can cluster related senses, resulting in more informative and interpretable representation.

Multiple markers may share similar meanings when connecting sentences. Thus, evaluating the performance of marker prediction simply on top1 accuracy is inappropriate. In Table 4, we demonstrated the results on ACC@k and observed that

[3]They also use the RobERTa model as a backbone.



(a) The cropped T-SNE visualization of discourse markers from the BASE PLM.



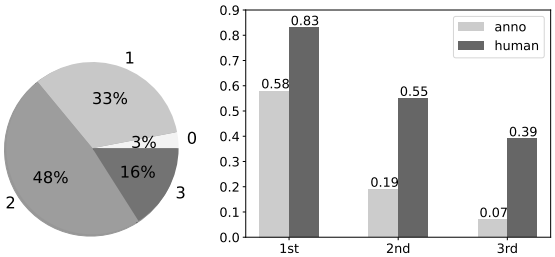
(b) The cropped T-SNE visualization of latent z from DMR.

Figure 4: The cropped T-SNE visualization of latent z from DMR. Each z is coupled with its top 3 markers from *z2m* mapping.

DMR(K=174) gets better performance against the model optimized by an MLE objective when k gets larger. We assume that it comes from the marker ambiguity. Our DMR models the ambiguity better, thus with any of the plausible markers easier to be observed in a larger range of predictions but more difficult as top1. To prove the marker ambiguity more directly, we randomly sample 50 examples to analyze their top5 predictions. The statistics show that over 80% of those predictions have plausible explanations. To conclude, considerable examples have multiple plausible markers thus ACC@k with larger k can better reflect the true performance on marker prediction, where DMR can beat the MLE-optimized model.

***z2m* Mapping** The latent space is not interpretable, but DMR has a transition matrix that outputs a distribution of markers, which reveals what a particular dimension may encode.

To analyze the latent space, we use ψ_{w2} (Eq. 5) as the corresponding embedding vectors and perform T-SNE visualization of the latent z, similar to what Discover (Sileo et al., 2019) does using the softmax weight at the final prediction layer. The complete T-SNE result can be found in Appendix B. What we observe is an emerging hierarchical pattern, in addition to proximity. That is, while synonymous markers are clustered as expected, semantically related clusters are often closer. Fig. 4b shows the top left corner of the T-SNE result. We can see that the temporal connectives and senses are located in the top left corner. According to their coupled markers, we can recover the semantic of these latent z: preceding (z_{27}), succeeding (z_{25}, z_{22}, z_{16}) and synchronous (z_{15}) form nearby but separated clusters.



(a) (b)

Figure 5: Human Evaluation. Figure (a) shows numbers of reasonable relations in top-3 predictions. Figure (b) shows the accuracy for each of the top-3 predictions evaluated by annotations or human, respectively.

For a comparison with connective-based prompting approaches, we also demonstrate the T-SNE visualization of marker representations from BASE in Fig. 4a. Unlike semantically aligned vector space of DMR, locality of markers in the space of BASE representation is determined by surface form of markers and shifted from their exact meaning. Marker representations of the model w/o latent z are closer because of similar lexical formats instead of underlying discourse.

From *z2m* mapping, we can take a step further to analyze the correlation between markers learned by DMR. Table 5 shows the top 2 corresponding clusters of three randomly sampled markers. We can observe correlations between markers like polysemy and synonym.

Understanding Entanglement Labeling discourse relations is challenging since some of them can correlate, and discern the subtleties can be challenging. For example, *List* strongly correlates with *Conjunction* and the two are hardly distinguishable.

DMR is trained to predict a *distribution* of markers, thus we expect its hidden state to capture the distribution of relations as well even when the multi-sense labels are scarce. We drew 100 random samples and ask two researchers to check whether each of the corresponding top-3 predictions is valid

s_1	s_2	markers	relations
Rather, they tend to have a set of two or three favorites	Sometimes, they'll choose Ragu spaghetti sauce	because_of_this therefore for_example for_instance	Contingency.Cause Expansion.Instantitation
It just makes healthy businesses subsidize unhealthy ones and gives each employer less incentive to keep his workers healthy	the HIAA is working on a proposal to establish a privately funded reinsurance mechanism to help cover small groups that can't get insurance without excluding certain employees	because_of_this conversely therefore in_contrast	Contingency.Cause Comparison.Contrast
The Hart-Scott filing is then reviewed and any antitrust concerns usually met	Typically, Hart-Scott is used now to give managers of target firms early news of a bid and a chance to use regulatory review as a delaying tactic	although though besides also	Comparison.Concession Expansion.Conjunction

Table 6: Case Study on Marker Ambiguity and Discourse Relation Entanglement.

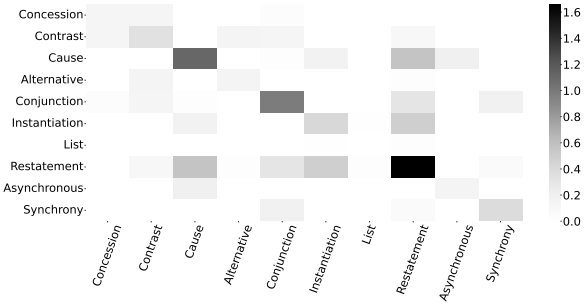


Figure 6: Confusion on Discourse Relations. We use entropy as the metric for filtering most confusing examples. We use the top-3 predictions of the 20 most confusing examples to show the entanglement between relations. We use accumulated $p(r_i) \cdot p(r_j)$ as weights for a pair of relations r_i, r_j. Note that implausible predictions are suppressed to ignore model errors.

and give a binary justification[4]. Fig. 5a shows that a considerable amount of 64% examples have two or more relations evaluated as reasonable in top-3 predictions, much higher than 2.3% multi-sense labels in PDTB2. This suggests that one way to improve upon the lack of multi-sense annotation is to use DMR to provide candidates for the annotators. For these samples, we also inspect annotator agreement in PDTB2 (Fig. 5b). While the trend is consistent with what DMR reports, it also validates again that the PTDB2 annotators under-labeled multi-senses.

To gain a deeper understanding of relation correlation, we rank the sentence pairs according to the entropy of relation prediction, a higher entropy suggests more model uncertainty, namely more confusion.

We use the top-3 predictions of the 20 highest entropy examples to demonstrate highly confusing discourse relations as shown in Fig. 6. The accumulated joint probability of paired relations on these examples is computed as weights in the confusion matrix. The statistics meet our expectation that there exist specific patterns of confusion. For example, asynchronous relations are correlated with causal relations, while another type of temporal relations, synchronous ones are correlated with conjunction. A complete list of these high entropy examples is listed in Appendix C.

To further prove DMR can learn diverse distribution even when multi-sense labels are scarce, we also evaluate our model on the DiscoGeM (Scholman et al., 2022), where each instance is annotated by 10 crowd workers. The distribution discrepancy is evaluated with cross entropy. Our model, trained solely on majority labels, achieved a cross entropy score of 1.81 against all labels. Notably, our model outperforms the BMGF model (1.86) under the same conditions and comes close to the performance of the BMGF model trained on multiple labels (1.79) (Yung et al., 2022). These results highlight the strength of our model in capturing multiple senses within the data.

To conclude, while we believe explicit relation labeling is still useful, it is incomplete without also specifying a distribution. As such, DMR's h_z or the distribution of markers are legitimate alternatives to model inter-sentence discourse.

[4]The annotators achieve a substantial agreement with a Kappa coefficient of 0.68.

Case Study on Specific Examples As a completion of the previous discussion on understanding entanglement in a macro perspective, we present a few examples in PDTB2 with markers and relations predicted by the DMR-based model. As demonstrated in Table 6, the identification of discourse relations relies on different emphasis of semantic pairs. Taking the first case as an example, the connection between "two or three favorities" and "Ragu spaghetti sauce" indicates the *Instantiation* relation while the connection between complete semantics of these two sentences results in *Cause*. Thanks to the probabilistic modeling of discourse information in DMR, the cases demonstrate entanglement among relations and ambiguity of markers well.

5 Conclusion

In this paper, we propose the distributed marker representation for modeling discourse based on the strong correlation between discourse markers and relations. We design the probabilistic model by introducing a latent variable for discourse senses. We use the EM algorithm to effectively optimize the framework. The study on our well-trained DMR model shows that the latent-included model can offer a meaningful semantic view of markers. Such semantic view significantly improves the performance of implicit discourse relation recognition. Further analysis of our model provides a better understanding of discourse relations and markers, especially the ambiguity and entanglement issues.

Limitation & Risks

In this paper, we bridge the gap between discourse markers and the underlying relations. We use distributed discourse markers to express discourse more informatively. However, learning DMR requires large-scale data on markers. Although it's potentially unlimited in corpus, the distribution and types of markers may affect the performance of DMR. Besides, the current solution proposed in this paper is limited to relations between adjacent sentences.

Our model can be potentially used for natural language commonsense inference and has the potential to be a component for large-scale commonsense acquisition in a new form. Potential risks include a possible bias on collected commonsense due to the data it relies on, which may be alleviated by introducing a voting-based selection mechanism on large-scale data.

References

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213, Austin, Texas. Association for Computational Linguistics.

Gillian Brown, Gillian D Brown, Gillian R Brown, George Yule, and Brown Gillian. 1983. *Discourse analysis*. Cambridge university press.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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

Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven C.H. Hoi, Caiming Xiong, Irwin King, and Michael Lyu. 2020. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2439–2449, Online. Association for Computational Linguistics.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

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.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Gabriel Murray. 2019. Discourse analysis and its applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–17.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1152–1158.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3830–3836.

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*.

Eric Malmi, Daniele Pighin, Sebastian Krause, and Mikhail Kozhevnikov. 2018. Automatic prediction of discourse connectives. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 368–375, USA. Association for Computational Linguistics.

Michael McCarthy, Matthiessen Christian, and Diana Slade. 2019. Discourse analysis. In *An introduction to applied linguistics*, pages 55–71. Routledge.

Linh The Nguyen, Linh Van Ngo, Khoat Than, and Thien Huu Nguyen. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4201–4207, Florence, Italy. Association for Computational Linguistics.

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.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone P. Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from CommonCrawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115.

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

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

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)*.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291.

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.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5790–5796.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(3):369–416.

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

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

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

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

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

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. *arXiv preprint arXiv:2210.07032*.

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 2010: Posters*, pages 1507–1514, Beijing, China. Coling 2010 Organizing Committee.

Train	Valid	Test
1566k	174k	174k

Table 7: Statistics of Discovery Dataset

Relations	Train	Valid	Test
Comp.Concession	180	15	17
Comp.Contrast	1566	166	128
Cont.Cause	3227	281	269
Cont.Pragmatic Cause	51	6	7
Expa.Alternative	146	10	9
Expa.Conjunction	2805	258	200
Expa.Instantiation	1061	106	118
Expa.List	330	9	12
Expa.Restatement	2376	260	211
Temp.Async	517	46	54
Temp.Sync	147	8	14
Total	12406	1165	1039

Table 8: Statistics of PDTB2 Dataset

A Implementation Details

We use Huggingface transformers (4.2.1) for the use of PLM backbones in our experiments. For optimization, we optimize the overall framework according to Algorithm 1. We train the model on Discovery for 3 epochs with the learning rate for ψ set to 3e-5 and the learning rate for ϕ set to 1e-2. The EM batchsize is set to 500 according to the trade-off between optimization efficiency and performance. The optimization requires around 40 hrs to converge in a Tesla-V100 GPU. For the experiments on PDTB2, we use them according to the LDC license for research purposes on discourse relation classification. The corresponding statistics of the two datasets are listed in Table 7 and Table 8.

B Visualization of the latent z

To obtain an intrinsic view of how well the connections between markers $\boldsymbol{m}$ and $\boldsymbol{z}$ can be learned in our DMR model. We draw a T-SNE 2-d visualization of $\boldsymbol{z}$'s representations in Fig. 7 with top-3 connectives of each z attached nearby. The representation vector for each z is extracted from ψ_{w2}. The results are interesting that we can observe not only the clustering of similar connectives as z, but also semantically related z closely located in the representation space.

C High Entropy Examples from Human Evaluation

For analysis of the entanglement among relations, we did a human evaluation on randomly extracted examples from PDTB2. To better understand the entanglement among relations, we further filter the 20 most confusing examples with entropy as a metric. The entanglement is shown as Fig.6 in Sec. 4.5. We list these examples in Table 9 for clarity.

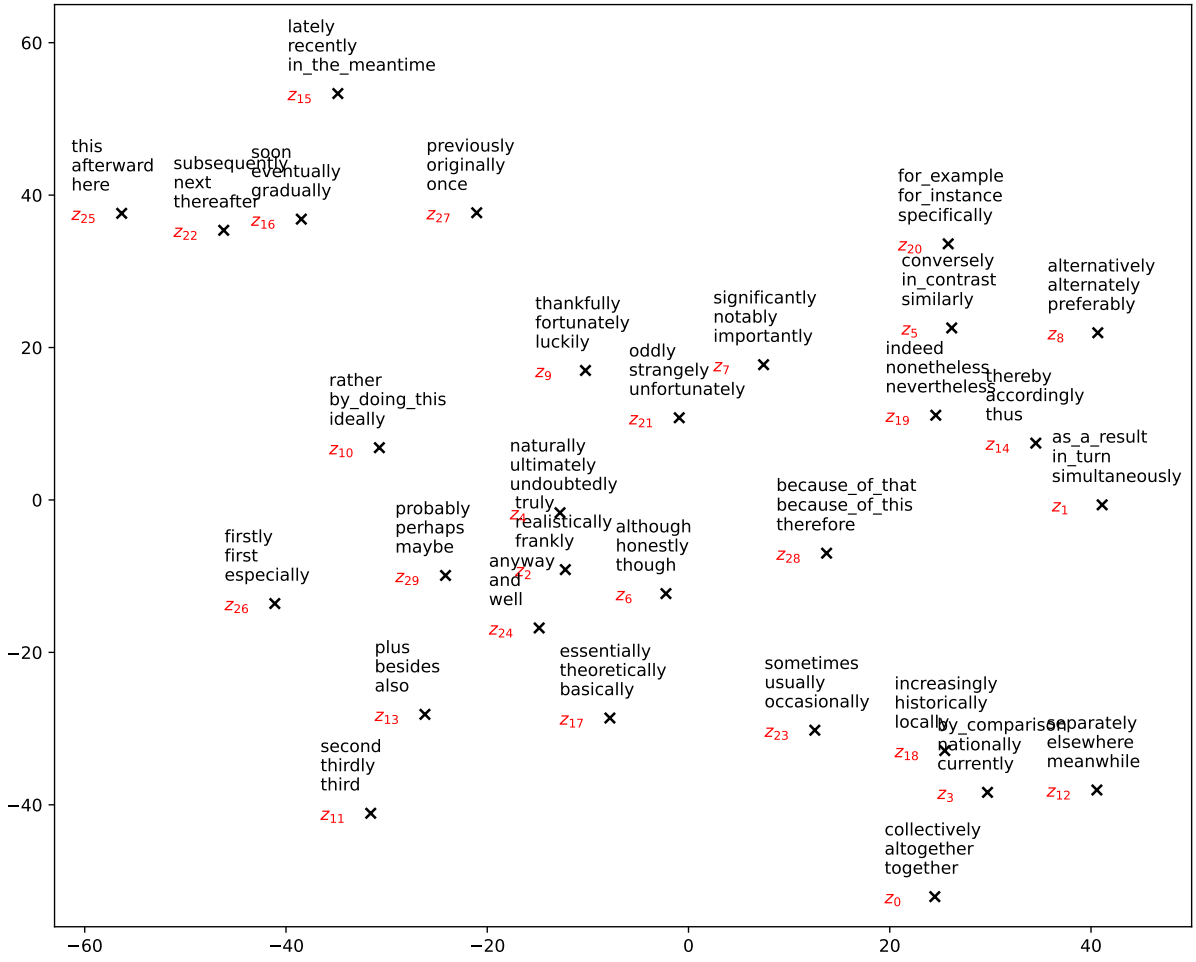


Figure 7: T-SNE Visualization of the Latent z. We draw the t-sne embeddings of each latent z in 2-d space with the well-trained ψ_{w2} as corresponding embedding vectors. While each z groups markers with similar meanings, we can also observe that related senses are clustered together. For example, temporal connectives and senses are located in the top left corner with preceding (z_{27}), succeeding (z_{25}, z_{22}, z_{16}), synchronous (z_{15}) ones separated. The existence of z helps to construct a hierarchical view of semantics between sentences.

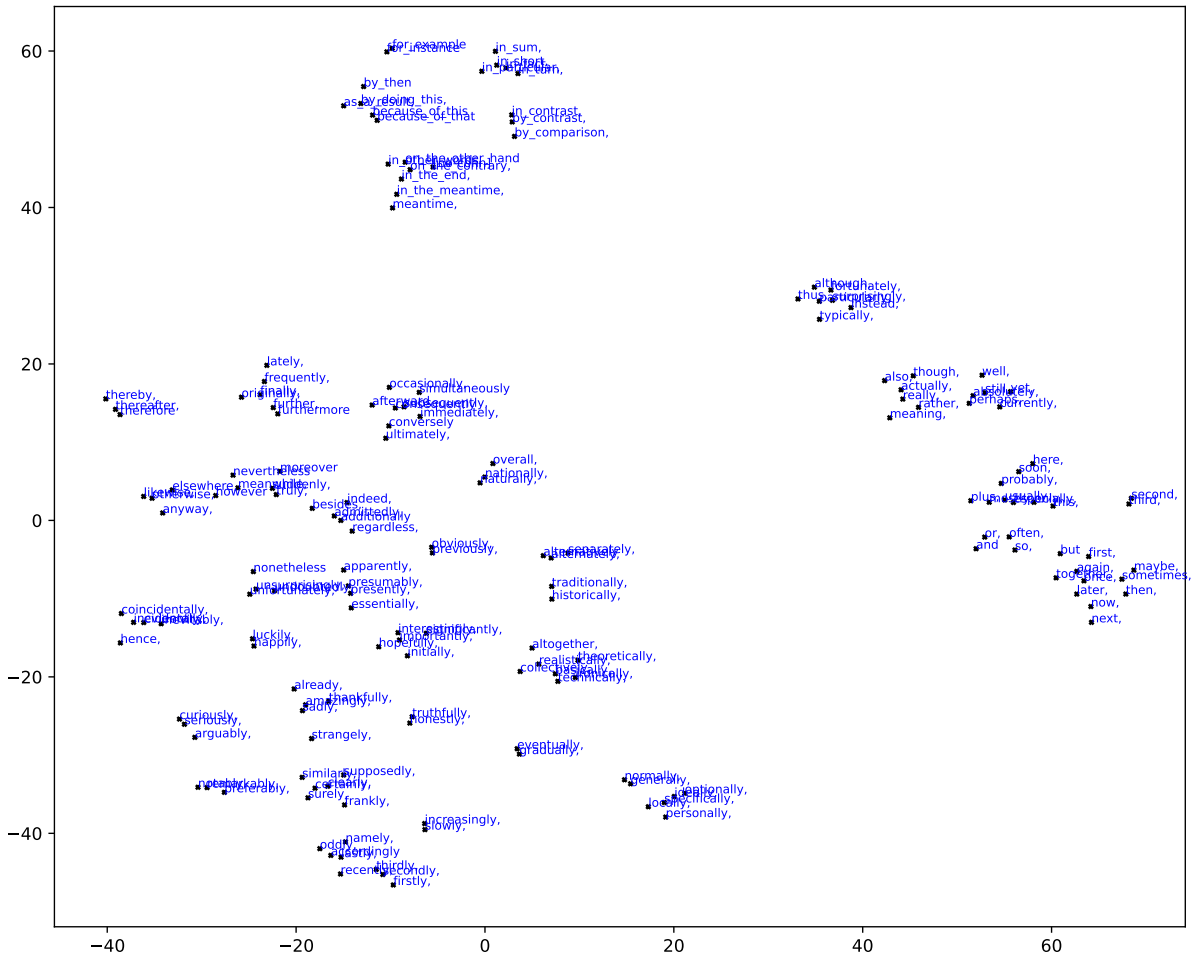


Figure 8: T-SNE Visualization of discourse markers from BASE. We draw the t-sne embeddings of each marker in 2-d space with averaged token representations of markers from BASE PLM. Comparing to the well-organized hierarchical view of latent senses in DMR, markers are not well-aligned to semantics in the representation space of BASE. It indicates the limitation of bridging markers and relations with a direct mapping.

s_1	s_2	1st-pred	2nd-pred	3rd-pred
Right away you notice the following things about a Philip Glass concert	It attracts people with funny hair	Instantiation 0.502	Restatement 0.449	List 0.014
There is a recognizable musical style here, but not a particular performance style	The music is not especially pianistic	Restatement 0.603	Conjunction 0.279	Instantiation 0.048
Numerous injuries were reported	Some buildings collapsed, gas and water lines ruptured and fires raged	Restatement 0.574	Instantiation 0.250	List 0.054
this comparison ignores the intensely claustrophobic nature of Mr. Glass's music	Its supposedly austere minimalism overlays a bombast that makes one yearn for the astringency of neoclassical Stravinsky, the genuinely radical minimalism of Berg and Webern, and what in retrospect even seems like concision in Mahler	Cause 0.579	Restatement 0.319	Instantiation 0.061
The issue exploded this year after a Federal Bureau of Investigation operation led to charges of widespread trading abuses at the Chicago Board of Trade and Chicago Mercantile Exchange	While not specifically mentioned in the FBI charges, dual trading became a focus of attempts to tighten industry regulations	Cause 0.504	Asynchronous 0.400	Conjunction 0.045
A menu by phone could let you decide, 'I'm interested in just the beginning of story No. 1, and I want story No. 2 in depth	You'll start to see shows where viewers program the program	Cause 0.634	Conjunction 0.188	Asynchronous 0.116
His hands sit farther apart on the keyboard.Seventh chords make you feel as though he may break into a (very slow) improvisatory riff	The chords modulate	Cause 0.604	Conjunction 0.266	Restatement 0.082
His more is always less	Far from being minimalist, the music unabatingly torments us with apparent novelties not so cleverly disguised in the simplicities of 4/4 time, octave intervals, and ragtime or gospel chord progressions	Cause 0.456	Restatement 0.433	Instantiation 0.052
It requires that "discharges of pollutants" into the "waters of the United States" be authorized by permits that reflect the effluent limitations developed under section 301	Whatever may be the problems with this system, it scarcely reflects "zero risk" or "zero discharge	Contrast 0.484	Cause 0.387	Concession 0.072
The study, by the CFTC's division of economic analysis, shows that "a trade is a trade	Whether a trade is done on a dual or non-dual basis doesn't seem to have much economic impact	Restatement 0.560	Conjunction 0.302	Cause 0.095
Currently in the middle of a four-week, 20-city tour as a solo pianist, Mr. Glass has left behind his synthesizers, equipment and collaborators in favor of going it alone	He sits down at the piano and plays	Restatement 0.357	Synchrony 0.188	Asynchronous 0.115
For the nine months, Honeywell reported earnings of $212.1 million, or $4.92 a share, compared with earnings of $47.9 million, or $1.13 a share, a year earlier	Sales declined slightly to $5.17 billion	Conjunction 0.541	Contrast 0.319	Synchrony 0.109
The Bush administration is seeking an understanding with Congress to ease restrictions on U.S. involvement in foreign coups that might result in the death of a country's leader	that while Bush wouldn't alter a long-standing ban on such involvement, "there's a clarification needed" on its interpretation	Restatement 0.465	Conjunction 0.403	Cause 0.094

s_1	s_2	1st-pred	2nd-pred	3rd-pred
With "Planet News Mr. Glass gets going	His hands sit farther apart on the keyboard	Synchrony 0.503	Asynchronous 0.202	Cause 0.147
The Clean Water Act contains no "legal standard" of zero discharge	It requires that "discharges of pollutants" into the "waters of the United States" be authorized by permits that reflect the effluent limitations developed under section 301	Alternative 0.395	Contrast 0.386	Restatement 0.096
Libyan leader Gadhafi met with Egypt's President Mubarak, and the two officials pledged to respect each other's laws, security and stability	They stopped short of resuming diplomatic ties, severed in 1979	Contrast 0.379	Concession 0.373	Conjunction 0.129
His hands sit farther apart on the keyboard.Seventh chords make you feel as though he may break into a (very slow) improvisatory riff.The chords modulate, but there is little filigree even though his fingers begin to wander over more of the keys	Contrasts predictably accumulate	Conjunction 0.445	Synchrony 0.303	List 0.181
NBC has been able to charge premium rates for this ad time	but to be about 40% above regular daytime rates	Conjunction 0.409	Restatement 0.338	Contrast 0.224
Mr. Glass looks and sounds more like a shaggy poet describing his work than a classical pianist playing a recital	The piano compositions are relentlessly tonal (therefore unthreatening), unvaryingly rhythmic (therefore soporific), and unflaggingly harmonious but unmelodic (therefore both pretty and unconventional	Cause 0.380	Instantiation 0.323	Restatement 0.241
It attracts people with funny hair	Whoever constitute the local Left Bank come out in force, dressed in black	Cause 0.369	Asynchronous 0.331	Conjunction 0.260

Table 9: High Entropy Examples of Model Inference on Implicit Discourse Relation Classification

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After Section 5

☑ A2. Did you discuss any potential risks of your work?
After Section 5

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

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

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4, Appendix A.

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

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☒ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4, Appendix A.

C ☑ Did you run computational experiments?

Section 4.

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

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4, Appendix A.

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

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

Left blank.

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

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

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

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

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