Improving Implicit Discourse Relation Recognition with Semantics Confrontation

Mingyang Cai[*1], **Zhen Yang**[*1], **Ping Jian**[†1,2]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Institute of Technology, Beijing, China
cai.my@foxmail.com, {bityangzhen, pjian}@bit.edu.cn

Abstract

Implicit Discourse Relation Recognition (IDRR), which infers discourse logical relations without explicit connectives, is one of the most challenging tasks in natural language processing (NLP). Recently, pre-trained language models (PLMs) have yielded impressive results across numerous NLP tasks, but their performance still remains unsatisfactory in IDRR. We argue that prior studies have not fully harnessed the potential of PLMs, thereby resulting in a mixture of logical semantics, which determine the logical relations between discourse arguments, and general semantics, which encapsulate the non-logical contextual aspects (detailed in Sec.1). Such a mixture would inevitably compromise the logic reasoning ability of PLMs. Therefore, we propose a novel method that trains the PLMs through two semantics enhancers to implicitly differentiate logical and general semantics, ultimately achieving logical semantics enhancement. Due to the characteristic of PLM in word representation learning, these two semantics enhancers will inherently confront with each other, facilitating an augmentation of logical semantics by disentangling them from general semantics. The experimental results on PDTB 2.0 dataset show that the confrontation approach exceeds our baseline by 3.81% F1 score, and the effectiveness of the semantics confrontation method is validated by comprehensive ablation experiments.

Keywords: Implicit discourse relation recognition, logical semantics enhancement, semantics confrontation

1. Introduction

Discourse relation recognition, aiming to identify the logical relation between two arguments (sentences or clauses), is a crucial task in discourse parsing which can benefit many downstream tasks in natural language processing (NLP), such as machine translation (Guzmán et al., 2014; Meyer and Popescu-Belis, 2012; Meyer and Webber, 2013) , text summarization (Gerani et al., 2014; Yoshida et al., 2014) and question answering (Jansen et al., 2014; Verberne et al., 2007). Discourse relation recognition encompasses two distinct paradigms: explicit and implicit. In explicit discourse relation recognition (EDRR) task, there are connectives (e.g., *because*, *so*) between the two discourse arguments, offering valuable prompts for the models to reasoning the semantic relations (as shown in **Example 1**). In contrast, implicit discourse relation recognition (IDRR) task lacks such connectives, necessitating models to extract logical information directly from the content of the arguments (as shown in **Example 2**).

In fact, EDRR has already demonstrated the remarkable effectiveness by utilizing explicit connectives. Pitler et al. (2008) have previously achieved a notably high accuracy of 93.09% in EDRR, while Zhou et al. (2010) demonstrated that a substantial F1 score of 91.8% can be attained only by pairing connectives with their corresponding discourse relations (e.g. *Reason*, *Result*). Nonetheless, IDRR is still an exceptionally challenging task due to the absence of the connectives, which poses greater challenges to the semantic understanding and logic reasoning ability of the models. Remarkably, even state-of-the-art methods can only attain an accuracy around 70%.

Example 1 (Explicit)

Arg1:We're offering this plan now
Conn:[Because]
Arg2:we feel it's the right time
Relation Sense:(*Contingency.Cause.Reason*)

Example 2 (Implicit)

Arg1:Living there for six years was really scary
Arg2:The ghosts of the past are everywhere
Relation Sense:(*Contingency.Cause.Reason*)

Recently, pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), have been widely adopted in many tasks of NLP, including IDRR (Liu et al., 2020; Wu et al., 2022). However, PLMs did not improve IDRR significantly. We argue that the primary reason is that the implicit relation recognition task imposes greater demands on text understanding and with the data-sparse nature of IDRR,

*Equal contribution
†Corresponding author

LREC-COLING 2024, pages 8828–8839
20-25 May, 2024.

Type	arg_1	arg_2
Original	Living there for six years was really scary	(because) The ghosts of the past are everywhere
Same logic	Living there was scary	(because) The ghosts are everywhere
Diff logic	Living there for six years was really scary	(although) The ghosts of the past are gone

Table 1: General and logical semantics of arguments

it inevitably presents formidable challenges for the models to achieve better performance.

To make it clear, firstly, we assume that any text unit consists of more than one aspect of semantics. As for IDRR, a robust discourse relation recognition model should be able to pronouncedly disentangle the logical semantics from logical-independent semantics (distinguished as "general semantics"). Here, the logical semantics play a pivotal role in the identification of logical relations between arguments, determining which specific relation they belong to. Conversely, the general semantics encapsulate all the non-logical contents, bearing little influence on logical relation reasoning.

Intuitively, the words contributing to logical semantics (denoted as logical words) are typically fewer in number compared to those contributing to general semantics (denoted as general words). As an example shown in Table 1, the logical relation is totally different when replacing just one word (*everywhere* $\rightarrow$ *gone*), which demonstrates a significantly shift in logical semantics. In contrast, the relation sense of the arguments still remains consistent when altering many other words. In other words, the distribution of logical words is more sparse compared to general words.

Meanwhile, a majority of PLMs employ mask language modeling (MLM) as their pre-training task. Given that the distribution of logical words are more sparse than general ones, the masking process mainly involves general words, which results in a predominant focus on learning the representation of general semantics, mixing with limited logical semantics. To verify this assertion, we obtained the embeddings of these sentences in Table 1 by putting them into a RoBERTa language model that had been fine-tuned using PDTB 2.0 dataset (Prasad et al., 2008). The distribution of these embeddings is visually depicted in Figure 1. As observed, sentences exhibiting different logical relations are found to be even closer to the original sentences than those sharing the same logical relations, which suggests that within the representations learned by PLMs, even after fine-tuning, there persists a mixing of general semantics and logical semantics.

In this paper, we propose a **S**emantics **C**onfrontation method for **I**mplicit **D**iscourse

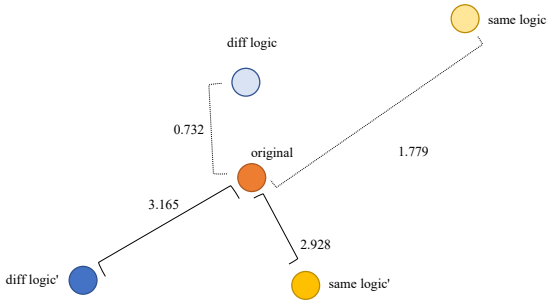

Figure 1: Distribution of the embeddings for different sentences in Table 1. The embedding of diff(same) logic is obtained by fine-tuned RoBERTa, while the embedding of diff(same) logic' is obtained by our model.

r**E**lation **R**ecognition (SCIDER) to disentangle the logical semantics from general semantics, thereby improving the performance of IDRR. Specifically, we firstly introduce two special tokens ([*general*] and [*logical*]) to represent general and logical semantics of arguments pairs, respectively. Subsequently, we train the pre-trained language model (RoBERTa) to learn the representation of both special tokens via different tasks (implemented by two **semantics enhancer**s). Inherently, these two tasks confront with each other, ultimately guiding the representations towards the respective spaces of general semantics and logical semantics. Notably, the disentangled logical semantics exhibit greater efficacy in clarifying logical relations. The experiments demonstrate that our approach exceeds our baseline by 3.81% F1 and 2.99% accuracy on PDTB 2.0[1].

It is noteworthy that the proposed approach can be applied not only to IDRR but also extended to any task requiring the disentanglement of the specific semantics (sentiment semantics, salient object features etc.) from general semantics (non-sentiment semantics, general image semantics etc.), such as sentiment classification, object detection and so on. So in essence, the **semantics enhancer** we proposed is a widely applicable framework with powerful transferability.

[1]Our code will be released at https://github.com/Young-Zhen/IDRR_SCIDER

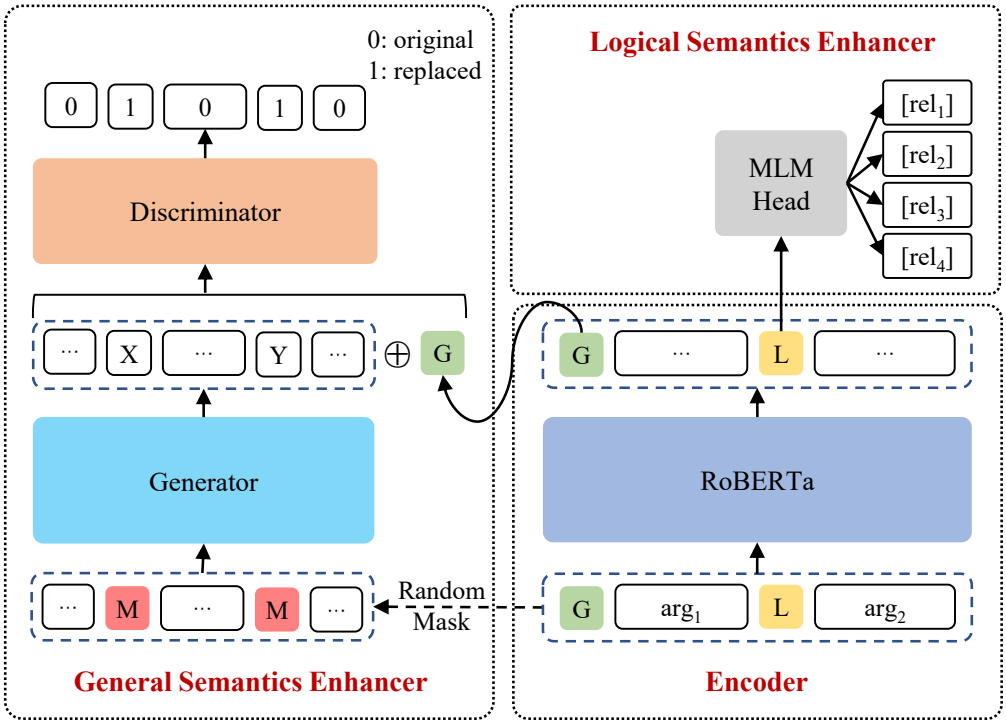

Figure 2: The model structure of our approach. The "G" represents "$[general]$" token, "L" represents "$[logical]$" token, "M" represents "$[MASK]$" token and $[rel_i]$ is manually designed token for each candidate logical relation, in order to be suitable for masked token predicting paradigm.

2. Related Work

Since the release of the PDTB corpus, numerous researchers have proposed various methodologies for IDRR task. In the early stages, the focus of researchers mainly revolved around manual feature engineering. For instance, Marcu and Echihabi (2002) derived word-pair features by calculating Cartesian products between all words in the two arguments and subsequently tallying their frequencies across each relation. Meanwhile, Varia et al. (2019) integrated both whole sentence and word-pair features to augment the features with global context information. Recognizing the potential data sparsity issue associated with word-pair features, Biran and McKeown (2013) addressed this concern by clustering semantically similar word pairs.

The emergence of neural networks has brought about significant enhancements in many NLP tasks including IDRR. Ji and Eisenstein (2015) employed two Recurrent Neural Networks (RNNs) to model argument semantics in a top-down manner and entity semantics in a bottom-up manner. Liu and Li (2016) harnessed the Long Short-Term Memory (LSTM) structure to mimic the repeatedly reading strategy, which can improve the comprehension of arguments. Qin et al. (2016) introduced stacked gated Convolutional Neural Networks (CNNs) to model arguments and perform classification. Inspired by the insight that the relation of arguments with connectives can be easily identified, Qin et al. (2017) adopted an adversarial approach to align the representations, ensuring that arguments without connectives exhibit similar representations to those with connectives. Dai and Huang (2018) posited that a broader context could aid argument understanding, motivating them to model the entire paragraph alongside the arguments themselves together."

In recent years, the deployment of large-scale pre-trained language models has significantly impacted various domains in NLP. Shi and Demberg (2019) adopted the BERT (Devlin et al., 2019) language model and retrained it using the next sentence prediction (NSP) task on PDTB. Similarly, Kishimoto et al. (2020) employed BERT as their encoder, but they conducted pre-train task on the discourse corpus before fine-tuning. Liu et al. (2020) proposed a method to learn the discourses' semantics in different level based on RoBERTa language model. On the other hand, Wu et al. (2022) explored the dependencies between discourse categories across different hierarchies and utilized a Graph Convolutional Network (GCN) to model these categories, they then combined the categories' and the arguments' RoBERTa embeddings to identify the discourse relations. Alternatively, Jiang et al. (2021) pursued a generative approach,

Type	Sentence
$input$	$[general]\ arg_1\ [logical]\ arg_2$
$label$	$[general]\ arg_1\ \ [rel_t]\ \ arg_2$

Table 2: Training data for logical semantics enhancer, wherein $[rel_t]$ represents the corresponding token of ground-truth logical relation between the argument pair arg_1 and arg_2.

wherein they designed a question specific to each discourse relation. Subsequently, they trained a T5 model not only to recognize the relation label but also to generate a target sentence that contains the meaning of the relations.

Notably, a trend has emerged in recent works towards the utilization of additional data. For instance, many works (Jiang et al., 2022; Chan et al., 2023b) employed the multi-level hierarchical information to enhance IDRR. Among them, GOLF (Jiang et al., 2022) exploited global and local hierarchies of senses through contrastive learning, while DiscoPrompt (Chan et al., 2023b) proposed a prompt-based method to predict the paths inside the hierarchical tree. More recently, PLSE (Wang et al., 2023) proposed a connective prediction based method which conducted pre-training with unannotated explicit data and then performed prompt-tuning with the implicit data.

Existing works have shown that the implicit discourse relation recognition is a significantly difficult task and the pre-trained language models play a crucial role in improving the performance. However, these approaches only utilize pre-trained models as the encoder, without further exploring the semantics representation generated by the PLMs.

3. Model

The overall structure of our model is illustrated in Figure 2. It delineates three key components [2]: the encoder, the logical semantics enhancer, and the general semantics enhancer. These components will be introduced individually in the subsequent sections.

3.1. Encoder

Following prevailing practices in this field, we utilize the pre-trained RoBERTa language model as the encoder. To enhance the model's ability in capturing the discourse semantics, we introduce two additional special tokens: "$[general]$" and "$[logical]$"., which represent general semantics and logical semantics, respectively. The input argument pair $(Arg1, Arg2)$ is reformulated as $([general], Arg1, [logical], Arg2)$. Intuitively, the "$[general]$" token is positioned before the first argument to capture the global, long-term, and "dense" general semantics. In contrast, the "$[logical]$" token is placed between the two arguments, in order to align more closely with the natural language expression where logical connectives typically emerge between two sentences.

Finally, the embeddings $\boldsymbol{e}_G$ for $[general]$ and $\boldsymbol{e}_L$ for $[logical]$ are learned through the encoder, and subsequently enhanced by two semantics enhancers.

3.2. Logical Semantics Enhancer

In the logical semantics enhancer, a manually designed relation token $[rel_i]$ is introduced for each candidate discourse relation. And a MLM head structure is utilized to predict the probability distribution of each word $token_i$ in vocabulary, based on the embedding of $[logical]$ token:

$$p(token_i|ctx) = \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{e}_L)}{\sum_{j=1}^{\textbf{vocab\_size}} \exp(\boldsymbol{w}_j^T \boldsymbol{e}_L)}, \quad (1)$$

where $\boldsymbol{e}_L$ is the embedding of $[logical]$ token, ctx is current context, $vocab\_size$ is the word count of language model's vocabulary, including the additional tokens, and $token_i$ includes the relation tokens $[rel_1], [rel_2], ..., [rel_n]$ and other normal tokens, wherein the $[rel_i]$ with the highest probability is served as the predicted relation label between two input arguments:

$$label = \underset{i}{\text{argmax}}\, p([rel_i]|ctx)\ (i = 1, 2, 3, ...). \quad (2)$$

In other words, the logical semantics enhancer regards $[logical]$ as a "$[MASK]$" token (as in BERT (Devlin et al., 2019)) and predicts the masked token $[rel_t]$ in MLM paradigm (the meaning of $[rel_t]$ and the specific data format for training are shown in Table 2). So the objective function for $[rel_t]$ prediction is as follows:

$$\mathcal{L}_{predict}([rel_t]|ctx) = \log p([rel_t]|ctx). \quad (3)$$

Besides, in order to enhance the comprehension of global semantics and further improve the model's performance in predicting the relation tokens $[rel_t]$, we also employ MLM as an auxiliary task. Specifically, we randomly mask some tokens, except $[logical]$, in the arguments, and then task the language model with predicting the masked tokens. The objective function of the MLM task can be described as Equation 4:

$$\mathcal{L}_{mlm}(\Pi|context) = \frac{1}{|\Pi|} \sum_{i=1}^{|\Pi|} \log p(\Pi_i|context), \quad (4)$$

[2] In Figure 2, it does not depict the Masked Language Modeling (MLM) task for input argument pairs, yet it still constitutes an essential part of our model to enhance the global understanding of discourse semantics for PLMs

where the Π is the collection of masked tokens, and $|\Pi|$ is the number of masked tokens.

Finally, as shown in Equation 5, the object function for training logical semantics enhancer is a sum of $\mathcal{L}_{predict}$ and $\mathcal{L}_{mlm}$:

$$\mathcal{L}_{logical} = \mathcal{L}_{predict} + \mathcal{L}_{mlm} \tag{5}$$

3.3. General Semantics Enhancer

As for the general semantics enhancer, we adopted the replaced token detection (RTD) task to gain expertise in capturing general semantics. It is inspired by the observation, as explained in Section 1, that such a specific MLM task is highly effective in learning general semantics. As illustrated in Table 3, tokens in the original $input$ sentences are randomly masked to get $input'$, then a generation model, with fixed parameters, predicts the masked token to obtain the $input''$. Meanwhile, the embeddings of $[general]$ token $\boldsymbol{e}_G$, obtained through the encoder, is concatenated with the words embeddings of $input''$, resulting in $\boldsymbol{e}_{RTD}$. Finally, a discrimination model is trained to check whether each token in $input''$ is the same as that in $input$. In order to improve the prediction accuracy of whether a word is replaced, the model need learn the $[general]$ token's representation with a more comprehensive grasp of all the general words, leading to a better understanding of general semantics. The loss function for general semantics enhancer is calculated as Equation 6:

$$\mathcal{L}_{general} = \sum_{t=1}^{|\boldsymbol{s}|} \Big(-\mathbb{1}\big(input''_t = input_t\big) \log D(\boldsymbol{e}_{RTD}) - \mathbb{1}\big(input''_t \neq input_t\big) \log\big(1 - D(\boldsymbol{e}_{RTD})\big) \Big) \tag{6}$$

where the $\mathbb{1}$ is indicator function, D is the discrimination model and $|s|$ is the length of the input sequence.

Type	Sentence			
$input$	The	ghosts	are	everywhere.
$input'$	The	$[MASK]$	$[MASK]$	everywhere .
$input''$	The	cats	are	everywhere.
$label$	0	1	0	0

Table 3: The different content of sentences after passing through the general semantics enhancer.

3.4. Semantics Confrontation

Through the two aforementioned semantics enhancers, the model has acquired the representation learning ability for logical and general semantics. In this section, we will explain how the semantics confrontation works. The two semantics enhancers achieve the logical and general semantics enhancement by relation token prediction (RTP) and replaced token detection (RTD), respectively. We leverage the inherent confrontation between the targets of RTP and RTD, thus training the two enhancers alternately to guide the representation of both tokens towards the logical and general semantics space, respectively. Ultimately, it facilitates an augmentation of logical semantics by disentangling the logical semantics from general semantics. Since we have adopted RoBERTa, a transformer-based PLM, as the encoder, due to the characteristics of self-attention in Transformer architecture, if one token's embedding is changed to fit a specific downstream task, it inevitably exerts an influence on the embeddings of other tokens. Therefore, the proposed method practically conducts implicit confrontation at the representation level, which enables the model to autonomously learn and allocate the two types of semantics, obviating the need to construct additional data for explicit adversarial learning.

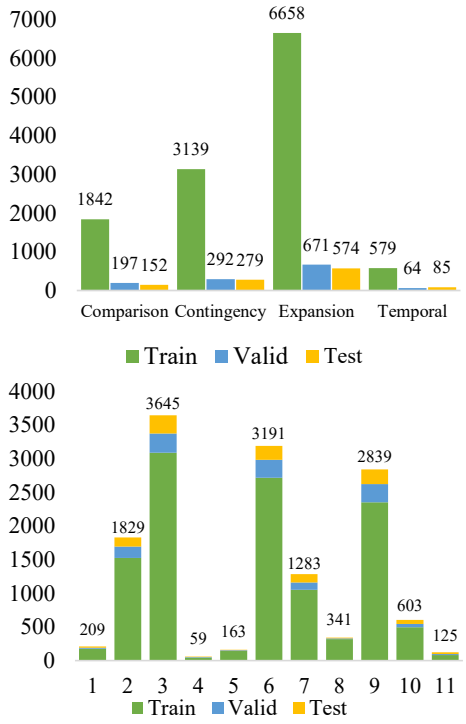

Figure 3: Data distribution categorized by the 4-way (up) and 11-way (bottom) classification of relation classes on PDTB 2.0 corpus

4. Experiments

4.1. Dataset

The PDTB 2.0 corpus is collected from more than 2,000 Wall Street Journal English articles (Prasad

et al., 2008), and divided into 24 different sections. There are both implicit and explicit discourse relation data in PDTB 2.0, which is classified into a hierarchical structure consisting of three levels. As shown in **Example 1** and **Example 2**, $Contingency$ belongs to the top-level sense, $Cause$ belongs to the second-level sense, and $Reason$ belongs to the third-level sense. We evaluate our method on the 4 top-level discourse classes and the 11 second-level discourse classes, consistent with prior studies (Long and Webber, 2022; Zhou et al., 2022a). Figure 3 provides an overview of the data distribution in PDTB 2.0. As depicted, the dataset comprises only about 14,000 instances of discourse data, making discourse relation recognition a **data-sparse** task. Furthermore, the dataset exhibits noticeable **data imbalance**, with the number of discourse pairs varying significantly across different relation categories. Following the predecessors (Ji and Eisenstein, 2015), we split sections 2-20, 0-1, and 21-22 as training, validation, and test sets respectively.

Another widely used dataset for IDRR is the CoNLL 2016 shared task (CoNLL16) (Xue et al., 2016). CoNLL16 merges several labels to annotate new relation senses and provides more abundant annotation. It consists of two test data denoted as CoNLL-Test (from PDTB section 23) and CoNLL-Blind (Wikinews texts).

4.2. Implementation Details

In the experiments, we utilize the RoBERTa-base model as the encoder. The logical semantics enhancer consists of a three-layer MLP with dimensions of 768, 768 and 4 for each layer, incorporating the Tanh activation function. In the auxiliary MLM task in logical semantics enhancer, we randomly mask 15% of the tokens empirically. As for the general semantics enhancer, we utilize a RoBERTa-base model as the generator, with parameters frozen. Here, the original input is randomly masked with a higher probability of 20% , which aims at accelerating the general words masking process and enhancing general semantics learning. Meanwhile, the discriminator comprises another RoBERTa-base model and an additional MLP, with the parameters learnable. We train the general semantics enhancer after logical semantics enhancer during one epoch. In other words, two semantics enhancers are trained alternately.

During the training process, we adopt AdamW (Loshchilov and Hutter, 2017) as the optimizer with an initial learning rate set to 3×10^{-6}, β_1 set to 0.9, and β_2 set to 0.999. The batch size is set to 16 for the concern of GPU memory occupation. We train our model for 20 epochs and select the checkpoints that yield the best performance on the validation dataset as the final model. All the experiments are performed on one 48GB NVIDIA A6000 GPU.

4.3. Baselines

To validate the effectiveness of the proposed method, we conduct comparative experiment with the most advanced baselines. Here we mainly introduce the baselines that have emerged within the past two years.

- **CG-T5** (Jiang et al., 2021) a generative approach, wherein a T5 model is trained to simultaneously recognize the logical relation and generate a sentence that contains the meaning of the relations.
- **CVAE-IDRR** (Dou et al., 2021) a CVAE-based method, which focuses on addressing the data-sparse problem in IDRR.
- **LDSGM** (Wu et al., 2022) a label dependence-aware sequence generation model, which exploits the dependence between hierarchically structured labels by a encoder-decoder structure.
- **PCP** (Zhou et al., 2022b) a prompt-based method, which utilizes the correlation between connectives and discourse relation using explicit connective prediction.
- **ChatGPT** (Chan et al., 2023a) a ChatGPT-based method leveraging ChatGPT with in-contex learning (ICL) prompt template.

Additionally, we fine-tune a RoBERTa-base model, without semantics confrontation and two special tokens, as our baseline model.

4.4. Experimental Results

Multi-way Classification Table 4 shows the performance of various models on the 4-way and 11-way classification on PDTB 2.0 and CoNLL16. The proposed method outperforms all the other models on top-level relation classification and surpasses our baseline model by a margin of 3.81% F1 score and 2.99% accuracy, demonstrating a remarkable improvement of the proposed method. It is worth noting that many second-level categories, as illustrated in Figure 3, are particularly **data-sparse**, which significantly hurts the accuracy of 11-way classification, including the proposed model. Compared to other works utilizing PLMs, such as BMGF-RoBERTa (Liu et al., 2020) and PCP (Zhou et al., 2022b), our method consistently shows noteworthy improvement in F1 score, indicating its efficacy in enhancing the PLMs's ability

Method	PDTB-4		PDTB-11	CoNLL-Test		CoNLL-Blind	
	F1	ACC	ACC	F1	ACC	F1	ACC
NNMA (Liu and Li, 2016)	46.29	57.17	-	-	-	-	-
Gshare (Lan et al., 2017)	47.80	57.39	-	-	39.40	-	40.12
Bi-LSTM-DU (Dai and Huang, 2018)	48.82	57.44	-	-	-	-	-
ELMo-C&E (Dai and Huang, 2019)	52.89	59.66	-	-	-	-	-
RWP-CNN (Varia et al., 2019)	50.20	59.13	-	-	39.39	-	39.36
KANN (Guo et al., 2020)	47.90	57.25	-	-	-	-	-
TransS (He et al., 2020)	51.24	59.94	-	-	-	-	-
BERT-HierMTN-CRF (Wu et al., 2020)	55.72	65.26	52.34	-	-	-	-
BERT-FT (Kishimoto et al., 2020)	58.48	65.26	54.32	-	-	-	-
BMGF-RoBERTa (Liu et al., 2020)	63.39	69.06	58.13	40.68	*57.26*	28.98	*55.19*
CG-T5 (Jiang et al., 2021)	57.18	65.54	53.13	-	-	-	-
CVAE-IDRR (Dou et al., 2021)	65.06	70.17	-	-	-	-	-
LDSGM (Wu et al., 2022)	63.73	71.18	60.33	-	-	-	-
PCP (Zhou et al., 2022b)	64.95	70.84	**60.54**	*33.27*	*55.48*	*26.00*	*50.99*
ChatGPT (Chan et al., 2023a)	36.11	44.18	24.54	-	-	-	-
Our Baseline Model	63.19	69.12	59.34	39.32	52.21	27.72	49.18
Ours (SCIDER)	**67.00**	**72.11**	59.62	**46.69**	**58.06**	**36.15**	**56.47**

Table 4: The macro-averaged F1 score (%) and accuracy (ACC) (%) of our model and previous works on PDTB 2.0 and CoNLL16. Italics numbers indicate the reproduced results from (Chan et al., 2023b). Bold numbers correspond to the best results, whereas underlined numbers correspond to the second best.

to reason the logical relation. Recently, large language models (LLMs), represented by ChatGPT, have demonstrated extraordinary ability across a spectrum of tasks, and Chan et al. (2023a) reported the performance of ChatGPT on IDRR task. Notably, ChatGPT's performance lags far behind other methods, displaying an inferiority of approximately 30% F1 and 28% accuracy compared to our approach, which indicates ChatGPT's limited ability of logical relation understanding without carefully designed prompt. Considering the outstanding performance of ChatGPT on many other tasks, there remains an evident and pressing need for further research on IDRR within the NLP community. Additionally, some recent studies, such as PLSE (Wang et al., 2023), GOLF (Jiang et al., 2022), and DiscoPrompt (Chan et al., 2023b), achieve good results by utilizing extra data, e.g., large-scale unannotated utterances with explicit connectives or annotated relation hierarchy structure. However, such methods exhibit an excessive reliance on additional data while our model can achieve remarkable performance based on discourse itself, which shows fertile avenues for future research explorations.

Binary Classification Table 5 presents the F1 score achieved by our model for each relation category of the 4 top-level classes. Notably, due to the limited data available for the "COMPARISON", "CONTINGENCY", and "TEMPORAL" categories, several models exhibit subpar performance. However, the proposed method consistently shows superior performance in the three categories compared to previous models, which verifies the effectiveness of our approach in mitigating the impact of **data imbalance** problem on the PDTB 2.0 dataset. The primary reason could be the fully data utilization facilitated by the general semantics enhancer, which is equivalent to harnessing the data twice beyond the logical semantics enhancer, thereby alleviating the imbalance in data distribution across different relation categories. However, as explained in Sec. 4.4, when the data is particularly sparse, it remains challenging and demanding to further improve the performance of IDRR.

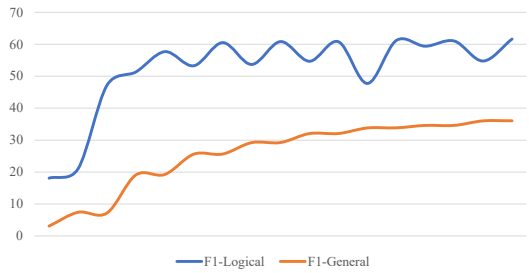

Figure 4: The trend of the F1 score during the training process. The results are obtained by conducting classification based on the embedding of $[logical]$ and $[general]$, respectively.

Visualization of Semantics Confrontation To investigate the influence of semantics confrontation, we conduct IDRR based on the representation

Model	COMPARISON	CONTINGENCY	EXPANSION	TEMPORAL
NNMA (Liu and Li, 2016)	36.70	54.48	70.43	38.84
Gshare (Lan et al., 2017)	40.73	58.96	72.47	38.50
Bi-LSTM-DU (Dai and Huang, 2018)	46.79	57.09	70.41	45.61
ELMo-C&E (Dai and Huang, 2019)	45.34	51.80	68.50	45.93
RWP-CNN (Varia et al., 2019)	44.10	56.02	72.11	44.41
TransS (He et al., 2020)	47.98	55.62	69.37	38.94
BMGF-RoBERTa (Liu et al., 2020)	59.44	60.98	77.66	50.26
KANN (Guo et al., 2020)	43.92	57.67	73.45	36.33
CG-T5 (Jiang et al., 2021)	55.40	57.04	74.76	41.54
CVAE-IDRR (Dou et al., 2021)	55.72	63.39	**80.34**	44.01
Ours (SCIDER)	**63.92**	**66.67**	78.12	**61.02**

Table 5: Binary classification results on PDTB for the 4 top-level classes of our baseline and previous works in terms of macro-averaged F1 score (%).

Method	F1	ACC
Classifying by $[logical]$		
w/o Conf	65.34	70.84
w/ Conf	66.31	71.76
Predicting $[rel_i]$		
w/o Conf	65.22	70.93
w/ Conf	**67.00**	**72.11**
replace $[general]$ *by* $\langle s \rangle$		
Classifying by $[logical]$	64.02	71.13
Predicting $[rel_i]$	64.85	68.36
train jointly	66.41	71.70
train alternately	**67.00**	**72.11**

Table 6: Ablation study on PDTB 2.0 in terms of multi-way classification, where Conf is the abbreviation of semantics confrontation.

of $[general]$ and $[logical]$, respectively. Specifically, for the acquired representations of both $[general]$ and $[logical]$ after each epoch, the model predicts the relation tokens through the MLM head in logical semantics enhancer. As shown in Figure 4, in the initial training epochs, the F1 score, obtained through the embedding of $[general]$, closely paralleled that of $[logical]$. But as training continued, a notable divergence emerged: $[logical]$ exhibited remarkable improvement while $[general]$ performs poorly, which indicates a pronounced disentanglement between logical and general semantics.

4.5. Ablation Study

4.5.1. Logical Semantics Enhancer and Semantics Confrontation

In the logical semantic enhancer, we introduce relation tokens $[rel_i], i = 1, 2, 3, ...$ for each discourse relation type, and conduct classification by predicting arguments' corresponding relation token $[rel_i]$ via a MLM head structure on top of $[logical]$'s embedding (denoted as "*predicting* $[rel_i]$"). However, another alternative method is using a MLP layer and getting the probabilities of each discourse relation instead (denoted as "*classifying by* $[logical]$"), as formulated in Equation 7:

$$p(relation = i) = \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{e}_L)}{\sum_{j=1}^{\text{num\_cls}} \exp(\boldsymbol{w}_j^T \boldsymbol{e}_L)}, \quad (7)$$

where $\boldsymbol{e}_L$ indicates special token $[logical]$'s embedding, and num_cls is the number of discourse relation types.

In addition, we also conducted an investigation into the impact of semantics confrontation on both classification strategies. Specifically, for the two strategies above, we trained our model with and without the general semantics enhancer, respectively. Table 6 shows the experimental results, in which the "w/o Conf" means without semantics confrontation. According to the experimental results, without semantics confrontation, directly classifying within 4 relation classes exhibits comparable performance with predicting $[rel_i]$ approach. Nevertheless, when conducting semantics confrontation, though both approaches exhibit improvement, the method of predicting rel_i outperforms the other, which demonstrates the importance and indispensability of MLM task in logical semantics enhancer.

4.5.2. General Semantics Representation

In the general semantic enhancer, we employ an additional special token $[general]$ to represent the general semantic of current discourse arguments. But when adopting RoBERTa, researchers usually use the existing special token $\langle s \rangle$, which represents the start of a sentence, for classification or other tasks. To explore the effect of additionally introduced special token $[general]$, we conduct experiment to compare its performance with utilizing the existing $\langle s \rangle$ token as the representation of general semantics. As the results shown in Table 6, the performance decreases across all settings when replacing $[general]$ with $\langle s \rangle$. The reason

could be that the representation of $\langle s \rangle$ contains lots of information unrelated to general semantics, consequently introducing interference with replaced token detection.

4.5.3. Train Semantics Enhancers Alternately or Jointly?

During the main experiments, the two semantics enhancers are trained alternately (as detailed in Section 4.2). In this section, we also seek to investigate the effect of joint training. In the case of joint training, the losses from both the logical and general semantics enhancers are combined, and we define the total loss, $\mathcal{L}_{joint}$, as a weighted sum of $\mathcal{L}_{logical}$ and $\mathcal{L}_{general}$:

$$\mathcal{L}_{joint} = \omega_l \mathcal{L}_{logical} + \omega_g \mathcal{L}_{general}, \quad (8)$$

where ω_l and ω_g are set to 5.0 and 1.0, empirically. As shown in Table 6, the joint learning shows inferior performance compared to alternate training, which indicates the merit of separately considering general and logical semantics enhancement during training. It demonstrates that the do-one-then-another paradigm of alternate training performs well in preventing the model from conflating logical and general semantics. The reason could be that when training jointly, the losses from both general and logical semantics enhancers simultaneously affect the encoder, inevitably causing the model conflating logical and general semantics. Correspondingly, the final learned representations of both $[general]$ and $[logical]$ are likely to be an average of the two semantics, which hurts the disentanglement of logical and general semantics.

5. Conclusion

In this paper, we argue that the representation of the arguments, learned by the PLMs, contains many aspects of semantics, wherein the logical semantics determines the discourse relation recognition. So we propose a novel semantics confrontation method to improve the performance of PLMs on IDRR. Our approach sufficiently disentangles the logical semantics from general semantics by introducing two special tokens ($[general]$ and $[logical]$) and implicitly guiding the representation of both tokens towards the semantics space of general and logical semantics, respectively. Experimental results indicate that our approach outperforms the concurrent methods for both 4-way and 11-way classification on PDTB 2.0 dataset.

Limitations

In the proposed work, we only consider the arguments' representation learned by PLMs as logical and general semantics, which may not cover all aspects of semantics. A fine-grained division may yield further improvement. Additionally, the semantics confrontation in our method is achieved implicitly, whereas explicit and carefully designed constraints may distinguish the logical and general semantics with a larger margin.

Ethics Statement

We note that this work mainly focuses on advancing technical aspects and conducting model evaluations, and the IDRR task is particularly data-sparse. Thus, we did not conduct additional aggressive data cleaning approach on the dataset, except those already applied to obtain the dataset. The text data utilized in our research might encompass elements of bias, toxicity, or unfairness. These aspects, although noteworthy, fall beyond the primary scope of our research, and we have not delved into them in specific detail. Apart from this, we have not identified any other potential risks.

Acknowledgement

The authors would like to thank the organizers of LREC-COLING 2024 and the reviewers for their helpful suggestions. This work is supported by the grants from the National Natural Science Foundation of China (No. 62172044).

Bibliographical References

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria. Association for Computational Linguistics.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023a. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *arXiv preprint arXiv:2305.03973*.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a

paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

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

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Fengyu Guo, Ruifang He, Jianwu Dang, and Jian Wang. 2020. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7822–7829.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

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*, pages 139–148, Online. Association for Computational Linguistics.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland. Association for Computational Linguistics.

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

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

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. *arXiv preprint arXiv:2211.13873*.

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 Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

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

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

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1233, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France. Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2263–2270, Austin, Texas. Association for Computational Linguistics.

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

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, Hong Kong, China. Association for Computational Linguistics.

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.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 735–736, New York, NY, USA. Association for Computing Machinery.

Chenxu Wang, Ping Jian, and Mu Huang. 2023. Prompt-based logical semantics enhancement for implicit discourse relation recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 687–699. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11486–11494.

Changxing Wu, Chaowen Hu, Ruochen Li, Hongyu Lin, and Jinsong Su. 2020. Hierarchical multi-task learning with crf for implicit discourse re-

lation recognition. *Knowledge-Based Systems*, 195:105637.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, Doha, Qatar. Association for Computational Linguistics.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022a. Prompt consistency for zero-shot task generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2613–2626, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022b. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Language Resource References

Rashmi Prasad and Nikhil Dinesh and Alan Lee and Eleni Miltsakaki and Livio Robaldo and Aravind K. Joshi and Bonnie L. Webber. 2008. *The Penn Discourse TreeBank 2.0*. European Language Resources Association, ISLRN 488-589-036-315-2.

Nianwen Xue and Hwee Tou Ng and Sameer Pradhan and Attapol Rutherford and Bonnie L. Webber and Chuan Wang and Hongmin Wang. 2016. *CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing*. ACL.