Interactively-Propagative Attention Learning for Implicit Discourse Relation Recognition

Huibin Ruan† **Yu Hong**†* **Yang Xu**† **Zhen Huang**‡ **Guodong Zhou**† **Min Zhang**†
† School of Computer Science and Technology, Soochow University, 1 Shizi, Suzhou, CHN
‡ School of Computer Science, National University of Defense Technology, Changsha, CHN
†{hbr416, tianxianer, andreaxu41}@gmail.com;
‡huangzhen@nudt.edu.cn; †{gdzhou, mzhang}@suda.edu.cn

Abstract

We tackle implicit discourse relation recognition. Both self-attention and interactive-attention mechanisms have been applied for attention-aware representation learning, which improves the current discourse analysis models. To take advantages of the two attention mechanisms simultaneously, we develop a propagative attention learning model using a cross-coupled two-channel network. We experiment on Penn Discourse Treebank. The test results demonstrate that our model yields substantial improvements over the baselines (BiLSTM and BERT).

1 Introduction

Implicit Discourse Relation Recognition (IDRR) is required to determine the relationship between arguments, under the condition that there is lack of a connective signaling the relationship. An argument generally stands for a narrative sentence or clause. For example, the arguments (i.e., Arg1 and Arg2) in Figure 1 hold a `causal` relation, where the possible connective "*because*" is not given.

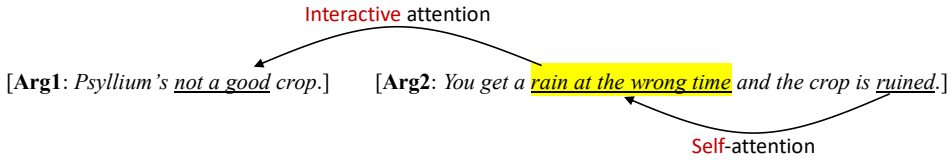

Figure 1: An example of causally-related arguments.

Since the time when IDRR was boiled down to a problem of discourse relation classification (Pitler et al., 2009; Lin et al., 2014), intense interest has been devoted to the study of argument representation and neural relation classification (Zhang et al., 2015; Liu et al., 2016; Chen et al., 2016; Lan et al., 2017; Bai and Zhao, 2018; Nguyen et al., 2019). Context-specific non-interactive attention mechanism (also referred to self-attention mechanism) (Lin et al., 2017) and companion-dependent interactive attention mechanism (Ma et al., 2017; Meng et al., 2016) have been used for enhancing the sentence-level embedding process. Both are proven effective in argument encoding (Guo et al., 2018; Liu and Li, 2016) as well as the perception of discourse relations. During encoding, the self-attention mechanism is able to highlight the latent information of attention-worthy words conditioned on local context (Note: we refer the attention-worthy words to the ones which play the dominant role in signaling discourse relations). By contrast, the interactive attention mechanism introduces external evidence into the identification of attention-worthy words, and similarly, highlighting their latent information.

So far, the two kinds of attention computations are performed separately. However, our survey shows that, in some cases, the context-specific self-attentive information can be inherited by the interaction-based attention computation. Let's consider the words "*rain at the wrong time*" in Arg2 in Figure 1. On the one hand, those words may catch the attention of the self-attention mechanism due to the occurrence

* Corresponding author: Yu Hong (tianxianer@gmail.com)

Proceedings of the 28th International Conference on Computational Linguistics, pages 3168–3178
Barcelona, Spain (Online), December 8-13, 2020

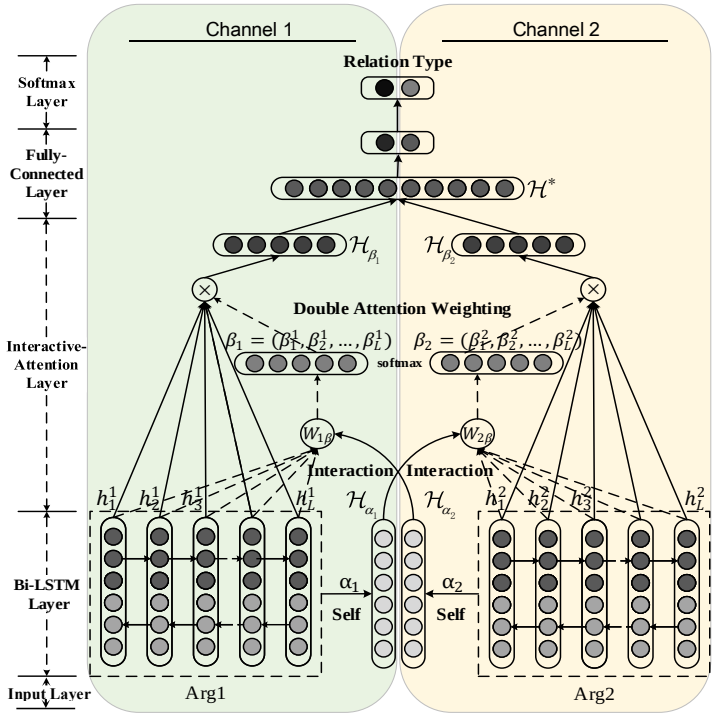

Figure 2: The bottom-up two-channel architecture of implicit discourse relation classification with bidirectional interactive attention propagation (GloVE embeddings are used in the input layer)

of the word "*ruined*" in context. On the other hand, they serve as reliable external evidence by which the interactive attention mechanism can recognize the crucial role of the words "*not a good*" in Arg1. This forms a multi-hop (2-hop) path along which a succession of self-to-interactive attention propagation may come into being (see the arrowed lines in Figure 1).

In order to model the attention-level continuity, we propose an Interactively-Propagative Attention Learning (IPAL) network. For a pair of arguments, IPAL deals with each of them independently, producing the self-attentive representation conditioned on the local context. Further, IPAL propagates the self-attentive information of one argument to the other. Using such information as the reliable external evidence, the interactive attention is computed between the arguments. Our experimental results show that IPAL outperforms both self and interactive attention mechanisms, and it has a competitive advantage when being integrated with BERT and multi-head self-attention mechanism.

The rest of the paper is organized as follows. Section 2 presents the architecture of IPAL-based implicit discourse relation classification, along with technical details (attention modeling, argument representation, loss estimation and training). In section 3, we show the experimental results obtained on PDTB v2.0. Besides, in this section, we compare IPAL with the state-of-the-art methods and verify its utility in perceiving the dominant attention-worthy words (by integrity verification and dominance examination). We overview the related work in section 4 and conclude the paper in section 5.

2 Approach (IPAL)

We show the overall framework of discourse relation classification model in Figure 2, where the pairwise argument analysis is performed. In the basic layer, BiLSTM is used to encode the arguments. In each bottom-up channel, self-attention is computed over the encoded argument, so as to produce the self-attentive representation. In the interactive layer, the resultant self-attentive representation in one channel will be delivered to the other channel. Consequently, in each channel, the interactive attention is computed using the delivered self-attentive representation as the external evidence. This allows the self-attentive information to be propagated to the generator of interactively-attentive representations, in a cross-channel manner. Eventually, the representations output by the two channels will be concatenated

to produce the final representation. Over the final representation, relation classification is conducted, where the dense and softmax layers constitute a multilayer perceptron.

2.1 BiLSTM Layer

The bidirectional recurrent neural network with LSTM (BiLSTM) (Schuster and Paliwal, 1997) is employed to encode the arguments. BiLSTM consists of forward and backward LSTM, which are used to capture the preceding and following information respectively. By BiLSTM, each word in an argument will be transformed into a forward hidden state $\overrightarrow{h_t} \in \mathbb{R}^{d_h}$ and a backward hidden state $\overleftarrow{h_t} \in \mathbb{R}^{d_h}$. We concatenate $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to form the synthetic hidden state $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. Accordingly, the inattentive sentence-level embeddings of the arguments (i.e., inattentive representations) can be represented as follows, where L is the maximum length of an argument:

$$\begin{cases} Arg1 : \mathcal{H}_1 \in \mathbb{R}^{L \times 2d_h} = (h_1^1, ..., h_L^1) \\ Arg2 : \mathcal{H}_2 \in \mathbb{R}^{L \times 2d_h} = (h_1^2, ..., h_L^2) \end{cases} \tag{1}$$

The entries of BiLSTM layer are constituted of pretrained GloVE word embeddings (Pennington et al., 2014). In our experiments, we additionally evaluate the effect of fine-tuned BERT (Devlin et al., 2019) which is deployed as the substitution of the GloVE based BiLSTM.

2.2 Classic Self-Attention Mechanism

For an argument, we first compute the self-attention vector $\alpha \in \mathbb{R}^L$ merely using intrinsic information in the argument itself. Lin et al. (2017)'s self-attention mechanism is used:

$$\alpha = \mathrm{softmax}(\check{W}_\alpha \tanh(W_\alpha \mathcal{H}^\top)) \tag{2}$$

where $W_\alpha \in \mathbb{R}^{d_a \times 2d_h}$ and $\check{W}_\alpha \in \mathbb{R}^{d_a}$ are learnable parameters, while d_a is a hyperparameter need to be tuned heuristically. On the basis, the self-attentive argument representation is computed as follows:

$$\mathcal{H}_\alpha = \alpha \mathcal{H} \tag{3}$$

2.3 Interactive Attention Propagation

We model the interaction between the inattentive and self-attentive argument representations. This enables the self-attentive information to be introduced into the computation of interactive attention. Suppose $\mathcal{H}$ is the inattentive representation of an argument, while $\mathcal{H}_\alpha$ is the self-attentive representation of the other argument, thus the interactive attention is computed with equation 4 (where, $\bar{\mathcal{H}}_\alpha$ is obtained by mean pooling: $\bar{\mathcal{H}}_\alpha = \sum_{i=1}^{L} \mathcal{H}_\alpha / L$, and $W_\beta \in \mathbb{R}^{2d_h \times 2d_h}$ and $b_\beta \in \mathbb{R}^L$ are learnable parameters).

$$\begin{cases} \beta = \mathrm{softmax}(\tanh(\mathcal{H} W_\beta \bar{\mathcal{H}}_\alpha^\top + b_\beta)) \\ \mathcal{H}_\beta = \beta \mathcal{H} \end{cases} \tag{4}$$

2.4 Bidirectional Attention Propagation

We carry out the interactive attention propagation for the two considered arguments in a bidirectional manner, from channel 1 to channel 2, and vice versa. Suppose that $\mathcal{H}_1$ and $\mathcal{H}_2$ stand for the inattentive representations of the arguments Arg1 and Arg2, and their self-attentive representations are $\mathcal{H}_{1_\alpha}$ and $\mathcal{H}_{2_\alpha}$ respectively (equation 3), thus the bidirectional interactive-attention vectors are computed as follows:

$$\begin{cases} \beta_1 = \mathrm{softmax}(\tanh(\mathcal{H}_1 W_{1_\beta} \bar{\mathcal{H}}_{2_\alpha}^\top + b_{1_\beta})) \\ \beta_2 = \mathrm{softmax}(\tanh(\mathcal{H}_2 W_{2_\beta} \bar{\mathcal{H}}_{1_\alpha}^\top + b_{2_\beta})) \end{cases} \tag{5}$$

Using the attention vectors β_1 and β_2, we compute the bidirectional interactive-attention representations $\mathcal{H}_{1_\beta}$ and $\mathcal{H}_{2_\beta}$ in equation 6, where $\mathcal{H}_{1_\beta}$ stands for the unidirectional interactive-attention representation of the argument Arg1, while $\mathcal{H}_{2_\beta}$ is that of Arg2:

$$\begin{cases} \mathcal{H}_{1_\beta} = \beta_1 \mathcal{H}_1 \\ \mathcal{H}_{2_\beta} = \beta_2 \mathcal{H}_2 \end{cases} \tag{6}$$

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

Table 1: Statistics of positive samples in the training, development and test sets (Note: If an argument pair serves as the positive sample of a certain relation type, it is the negative sample of other types).

2.5 Discourse Relation Classification

We concatenate $\mathcal{H}_{1_\beta}$ and $\mathcal{H}_{2_\beta}$ to form the final representation: $\mathcal{H}^* = [\mathcal{H}_{1_\beta}, \mathcal{H}_{2_\beta}] \in \mathbb{R}^{4d_h}$. We feed $\mathcal{H}^*$ into the multilayer perceptron to estimate the probability $\hat{y}_r$ that the arguments hold a relation r:

$$\begin{cases} \hat{y}_r = \text{softmax}(W\mathcal{H}^* + b) \\ r = \underset{r \in \mathcal{R}}{\arg\max}\ \hat{y}_r \end{cases} \tag{7}$$

where $W \in \mathbb{R}^{n \times 4d_h}$ and $b \in \mathbb{R}^n$ are trainable parameters, and $\mathcal{R}$ stands for the predefined set of discourse relation classes and n denotes the number of classes.

2.6 Training

We perform binary classification for each of the four types of PDTB relations (Comparison, Contingency, Expansion and Temporality), determining whether a pair of arguments holds a specific type of relation. Given a target relation type r, the set $\mathcal{R}$ comprises two (n=2) class labels— $r_\triangle$ and $r_\triangledown$— which respectively signal a positive sample (argument pair) which holds the target relation and a negative sample which doesn't hold the relation.

During training, we minimize the binary classification loss $\mathcal{L}$. The cost function is the cross-entropy of y_r and $\hat{y}_r$ for both the class labels $r_\triangle$ and $r_\triangledown$, where $y_{r_\triangle}, y_{r_\triangledown} \in \{0,1\}$ denotes the ground truth:

$$\mathcal{L} = -y_{r_\triangle} \log(\hat{y}_{r_\triangle}) - y_{r_\triangledown} \log(\hat{y}_{r_\triangledown}) \tag{8}$$

3 Experimentation

3.1 Datasets and Evaluation Metric

We follow the common practice to use section 02-20 of PDTB v2.0 (Prasad et al., 2008) as the training set, section 00-01 as the development set, and section 21-22 as the test set. Table 1 shows the statistics of instances in the sets. We use F1-score as the evaluation metric for binary discourse relation classification. Besides, in the discussion sections, P-value (Johnson, 1999) is taken as the evaluation metric for statistical significance, and NDCG@k (Järvelin and Kekäläinen, 2002) is employed for evaluating the integrity of attention-worthy words.

3.2 Hyperparameter Settings

We set two groups of hyperparameters in total, which correspond to different word embedding learning models (i.e., pretraining models): GloVE and BERT respectively.

When GloVE (Pennington et al., 2014) is used, we set the dimension of word embedding to 50 and the maximum length of argument 80 (L=80). During training, we set the mini-batch size to 32 and specify the dropout rate as 0.1 (Srivastava et al., 2014). Besides, each of LSTM units is of 50 dimensions (d_h=50) and the number of hidden states in the self-attention layer is set to 80 (d_a=80). We initialize the trainable parameters by randomly sampling in [-0.1, 0.1]. The learning rate for parameter updating is set to 1e-3.

In a separate experiment, we modify our model by integrating the pretrained BERT instead of GloVE, where BERT (Devlin et al., 2019) is fine-tuned. The dimension of each hidden state output by BERT is set

conditioned on **GloVE embeddings**	**COM**	**CON**	**EXP**	**TEM**
Bi-LSTM (Baseline 1)	33.33	50.75	67.84	29.81
+Self-Attention	35.43	52.26	68.47	35.99
+Multihead-Attention	33.40	53.19	70.68	31.42
+Interactive-Attention	36.67	52.25	69.22	33.69
+IPAL	**37.09**	**54.21**	70.86	**37.56**
+IPAL+Multihead	31.85	52.37	**71.46**	27.00
conditioned on **BERT embeddings**	**COM**	**CON**	**EXP**	**TEM**
Fine-tuned BERT (Baseline 2)	41.14	55.67	73.39	35.34
+Self-Attention	45.45	57.58	74.94	34.51
+Multihead-Attention	45.00	57.95	75.08	37.16
+Interactive-Attention	45.85	59.35	74.90	36.02
+IPAL	**46.88**	57.98	75.27	36.31
+IPAL+Multihead	46.75	**59.56**	**75.83**	**39.35**

Table 2: Test results for discourse relation classification in the ablation study

Conditioned on **GloVE**	BiLSTM	BiLSTM+Self	BiLSTM+Multihead	BiLSTM+Interactive
BiLSTM+IPAL	**0.013***	**0.001***	**0.067**	**0.035***
BiLSTM+Multihead+IPAL	0.441	0.214	0.131	0.176
Conditioned on **BERT**	BERT	BERT+Self	BERT+Multihead	BERT+Interactive
BERT+IPAL	0.040*	**0.037***	0.311	0.443
BERT+Multihead+IPAL	**0.004***	0.043*	**0.007***	**0.072**

Table 3: Resutls of statistical significance tests

to 768. We follow Devlin et al (2019) to reset the learning rate to 5e-5. The rest of the hyperparameters remain unchanged. The source code to reproduce the experiments will be made publicly available.

3.3 Main Results

We use BiLSTM and fine-tuned BERT as the baselines, which are respectively connected with a one-layer MLP for discourse relation classification. Though they are not integrated with any kind of attention mechanism. We evaluate the effects of different attention mechanisms when they are coupled with the baselines. The considered attention mechanisms include the classic self (Lin et al., 2017), multi-head self (Vaswani et al., 2017) and interactive versions (Chen et al., 2016; Bai and Zhao, 2018; Nguyen et al., 2019), as well as our bidirectional IPAL. Table 2 shows the test results. It can be observed that IPAL produces significant improvements over the baselines, and it outperforms other attention mechanisms.

In particular, the attention mechanisms obtain better performance when cooperating with the fine-tuned BERT. Compared to the classic self-attention mechanism, the multi-head self-attention mechanism is more compatible with BERT when dealing with higher-dimensional embeddings. Considering this fact, in a separate experiment, we take the multi-head version to form IPAL instead of the classic one. It can be observed that the updated IPAL achieves better performance (see the bottom row in Table 2).

3.4 Discussion 1: Statistical Significance Testing

We follow Johnson (1999) to use the sampling-based P-values for examining the significance. Johnson (1999) suggests that the ideal threshold of P-value is 0.05. It indicates that a system achieves significant improvements over others only if P-values are less than 0.05, otherwise insignificant. More importantly, it has been demonstrated that the smaller the P-value, the higher the significance (Dror et al., 2018). We calculate P-values by comparing the experimental results of IPAL and the updated version (Multihead self-attention+IPAL) with those of others. Similarly, we consider two scenarios in which the GloVE based BiLSTM and BERT respectively cooperate with IPAL. We show the results of significance tests in Table 3, where the P-values which are lower than the threshold are marked with the sign "$*$".

Method	COM	CON	EXP	TEM
Zhang et al. (2015)	33.22	52.04	69.59	30.54
Chen et al. (2016)	40.17	54.76	-	31.32
Qin et al. (2016)	41.55	57.32	71.50	35.43
Liu et al. (2016)	37.91	55.88	69.97	37.17
Liu and Li (2016)	36.70	54.48	70.43	38.84
Qin et al. (2017)	40.87	54.56	72.38	36.20
Lan et al. (2017)	40.73	58.96	72.47	38.50
Dai and Huang (2018)	46.79	57.09	70.41	**45.61**
Lei et al. (2018)	43.24	57.82	72.88	29.10
Guo et al. (2018)	40.35	56.81	72.11	38.65
Bai and Zhao (2018)	47.85	54.47	70.60	36.97
Nguyen et al. (2019)	**48.44**	56.84	73.66	38.60
He et al (2020)	47.98	55.62	69.37	38.94
IPAL+Multihead (BERT)	46.75	**59.56**	**75.83**	39.35

Table 4: Comparisons with the state of the art

The P-values listed in Table 3 demonstrate that IPAL yields statistically significant improvements over the baseline (BiLSTM), classic self-attention and interactive attention mechanisms when GloVE is used. On the other hand, when BERT is used, IPAL yields significant improvements over the baseline and classic self-attention mechanism. Besides, in this scenario, the updated IPAL produces significant improvements over most competitors except the interactive attention mechanism. In addition, it can be observed that, in the scenario where GloVE is used, IPAL achieves lower P-values than the updated version, but, on the contrary, most P-values the updated IPAL obtained are lower when BERT is used.

Considering the findings in the significance tests and the fact that the input BERT embeddings are of higher dimension than GloVE (768 versus 100 in our case), we recommend the following precautions towards the practical application of IPAL: first, an isolated IPAL applies more to the representation learning in the low-dimensional semantic space; second, if IPAL is intentionally used for the high-dimensional representation learning, it needs to be coupled with multi-perceptive semantic space transformation models, such as the multi-head self-attention mechanism in transformer (Vaswani et al., 2017).

3.5 Discussion 2: Compared to the State of the Art

There are a variety of advanced techniques have been developed, in which sophisticated networks were successfully constructed and reliable features were carefully exploited.

- For **attention learning**, Liu et al (2016) developed a multilayer attention mechanism. Chen et al (2016) integrated both the linear and non-linear interactions. Guo et al (2018) incorporated sparse learning into the interactive attention mechanism. Bai and Zhao (2018) used a feed forward network to model interactive attention and captured the effects on multi-grain linguistic features. Nguyen et al (2019) followed Bai and Zhao (2018)'s framework and conducted knowledge transferring.

- For **model design**, neural networks were mainly used, including the basic ones like CNN, RNN and LSTM (Zhang et al., 2015; Qin et al., 2016; Guo et al., 2018), and the variants such as CRN and CGNN (Chen et al., 2016; Qin et al., 2016), as well as adversarial (Qin et al., 2017) and multi-task learning models (Lan et al., 2017; Bai and Zhao, 2018; Nguyen et al., 2019).

- For **feature selection**, the embeddings of character, subword, word, sentence and sentence-pair levels (Bai and Zhao, 2018; Nguyen et al., 2019), paragraph-level relation continuity (Dai and Huang, 2018) and topic continuity (Lei et al., 2018) have been successfully applied in this area.

Table 4 shows the performance of the previous methods and ours. Compared to the state-of-the-art methods mentioned above, our IPAL is puny as it is isolated from the highly sophisticated learning architectures. More seriously, it has not yet utilized diverse features or other closely-related data resources. As

a result, the basic IPAL equipped with GloVE fails to achieve a competitive performance. Nevertheless, IPAL is "vest-pocket" and therefore can be easily assembled with other models to form a complicated system. As shown in Table 4, the updated version of IPAL which is coupled with BERT and multi-head self attention achieves the best performance for the contingent (CON) and expansive (EXP) relations. For the Temporal (TEM) relation, it ranks second behind Dai and Huang (2018)'s method, a neural network which models both paragraph-level argument dependency and discourse relation continuity. For the comparative (COM) relation, it ranks third behind the methods of Bai and Zhao (2018) and Nguyen et al. (2019), both of which conduct multi-task learning, and they either utilize the embeddings of multi-grain linguistic units, or develop knowledge transferring via relation and connective embeddings.

3.6 Discussion 3: Integrity versus Dominance

We tend to verify whether a neural attention model is able to identify as many attention-worthy words as possible in a pair of arguments. Therefore, in a separate experiment, we carry out the integrity verification. In addition, it is believed that the attention-worthy words would play a dominant role in signaling the argument relations, and thus they should catch more considerable attention than the ordinary words (i.e., the attention-unworthy words). Hence, we further examine the dominant effects of attention-worthy words. In this subsection, we first introduce the annotation of attention-worthy words; second, we present the NDCG-based integrity verification; finally, we show the average divergence of dominance between attention-worthy and ordinary words. The evidence provided by this case study will demonstrate that the dominant effects of attention-worthy words are more important for discourse relation perception.

- ***Labeling of Attention-worthy Words***

We conduct a case study on a set of 400 argument pairs. Two experienced masters who study on computational linguistics annotated 100 argument pairs for each relation type (All the four primary relation types mentioned in Table 1 are considered). The argument pairs are selected randomly. During the annotation process, if the annotators disagree with each other on a certain instance, they will be asked to debate until a decision is made, otherwise the instance will be replaced by a newly-selected substitute.

The attention-worthy words in the arguments are labeled as 1 (y_i=1), while the ordinary words are labeled as 0 (y_i=0). The following example shows the one-hot annotation results over two arguments.

	0	1	0	0	0
[**Arg1**: *It*	*began*	*to*	*turn*	*around.*]	

	0	1	0	0	0	0	0	0
[**Arg2**: *It*	*ended*	*with*	*a*	*gain*	*of*	*88.12*	*points.*]	

- ***NDCG-based Integrity Verification***

Normalized Discounted Cumulative Gain of top-k ranking results (NDCG@k) (Järvelin and Kekäläinen, 2002) is a widely used metric for evaluating learning-to-rank methods in the field of information retrieval. We regard the NDCG@k value as the gold standard and use it to evaluate the attention-based ranking results for words, verifying whether the attention-worthy words have been ranked ahead of the ordinary words. The higher the attention-worthy words are ranked (in terms of attention scores), the more of them can be perceived by a neural encoder. We name the investigation of such a ranking order as the "integrity verification" of the perceived attention-worthy words.

For an argument X which consists of L words, i.e., $X = \{x_1, x_2, ..., x_L\}$, and the annotation results $Y = \{y_1, y_2, ..., y_L\}$ on the words, we first use an attention mechanism to calculate the attention scores $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_L\}$ for all the words in X respectively. Relying on the attention scores, we rank the words in X in descending order. On the basis, we calculate the NDCG score for top-k words as follows:

$$NDCG@k = \frac{DCG_k}{IDCG_k} \quad DCG@k = \sum_{i=1}^{k_{\mathrm{PRE}}} \frac{y_i}{log(i+1)} \quad IDCG@k = \sum_{i=1}^{k_{\mathrm{REL}}} \frac{y_i}{log(i+1)} \tag{9}$$

where, k_{PRE} refers to the top-k words in the ranking list L_{PRE}, and k_{REL} is that in the ranking list L_{REL}. L_{PRE} is obtained conditioned on the attention scores $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_L\}$, while L_{REL} is obtained in terms of the annotation results $Y = \{y_1, y_2, ..., y_L\}$ (i.e., the real condition).

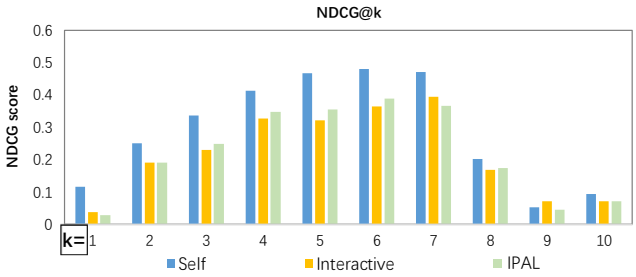

Figure 3: NDCG scores at different k that are obtained by different attention models

In equation 9, $DCG@k$ stands for the discounted cumulative gain of the top-k words in the predicted ranking list L_{PRE}. By contrast, $IDCG@k$ is an ideal discounted cumulative gain of the top-k words in the manually-crafted ranking list L_{REL}. Therefore, NDCG is actually a measure of agreement between man-made and automatically-generated ranking lists. In our case, the attention-worthy words have been labeled as 1 and, ideally, they should be ranked ahead of the ordinary words (labeled as 0). As a result, only if an attention mechanism assigns higher attention scores to all the attention-worthy words rather than the ordinary ones, it can achieve a higher NDCG score.

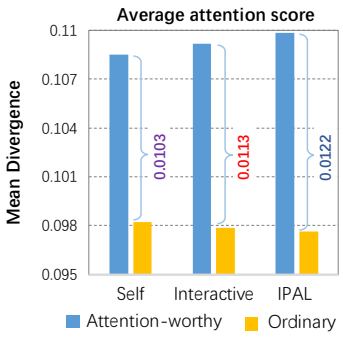

Figure 4: Mean divergences between attention-worthy and ordinary words

It can be found that when k is changed in a range from 1 to 10, IPAL obtains slightly higher NDCG@k scores than the interactive attention mechanism (except the cases when k equals to 1, 7 and 9). However, IPAL performs considerably worse for NDCG@k than the self-attention mechanism. The results demonstrate that, compared to IPAL, the self-attention mechanism is more capable of capturing attention-worthy words, with a lower missing rate. This raises the question about why IPAL outperforms the self-attention mechanism for argument-level relation classification (see Table 2). In order to explain the contradictory results, we went further to investigate the attention distribution and subsequently observed the noticeable effect of dominant attention-worthy words (see next subsection).

- ***Dominance Examination by Average Divergence***

We suppose that, if a word is assigned a higher attention weight, it plays a dominant role in signaling the relation of a pair of arguments. Thus, the divergence of dominant effects of words can be equivalently represented as the quantitative difference of attention weights assigned to them. In this case study, we investigate the attention weights which are assigned by different attention mechanisms to the ground-truth attention-worthy and ordinary words, and calculate the average divergences of dominant effects between the two kinds of words.

Figure 4 shows the mean attention weights along with the average divergences. It can be observed that the use of IPAL results in a greater divergence than the self-attention mechanism. In other words, by assigning much higher attention weights to the truly attention-worthy words rather than the ordinary ones, IPAL enables the attention-worthy words to exert more dominant effects on the perception and classification of discourse relations. Considering that IPAL does perform better than the self-attention mechanism for discourse relation classification, we suggest that improving dominant effects of some of attention-worthy words may be more important than the identification of all the attention-worthy words. In Figure 5, we give the examples in which words are highlighted with the predicted attention weights.

4 Related Work

In the earlier work, the traditional machine learning techniques (SVM and Bayes classifiers) were used for discourse relation classification. Meanwhile, a great interest has been devoted to the empirical find-

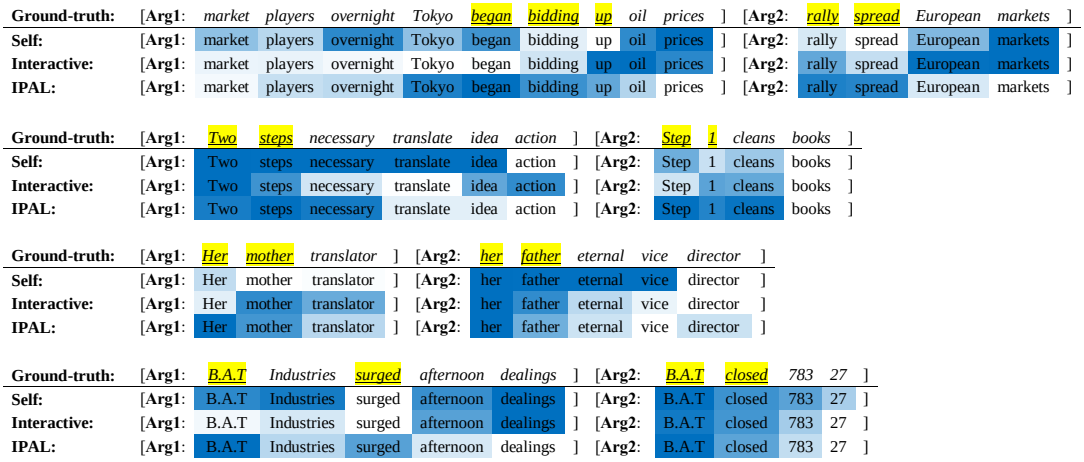

Figure 5: Examples of attention weight assignment (Ground-truth attention-worthy words are marked by yellow background, while the predicted ones by blue. A darker color denotes a higher attention weight.)

ings, as well as the exploitation of effective features. One of the most important feature engineering approaches uses interrelated word pairs as the reliable features (Marcu and Echihabi, 2002) since they imply semantic relationships. Hereafter, part-of-speech (POS) (Pitler et al., 2009), syntactic structures and dependencies (Lin et al., 2009) and semantic properties (Lei et al., 2018) were used as novel features.

Recently, neural networks have been widely studied for argument representation learning (Zhang et al., 2015), which is admitted to be the crucial issue for discourse relation recognition. Due to the capacity of generating low-dimensional continuous representations for arguments, RNNs with Bi-LSTM are used during encoding. Chen et al (2016) couple Bi-LSTM with a gated relevance model. Liu and Li (2016) use multi-layer attention computation over the output of Bi-LSTM. Meanwhile, Liu et al (2016) build a multi-task learning framework with Convolutional Neural Network (CNN) for argument encoding. By contrast, Lan et al (2017) integrate Bi-LSTM into the multi-task framework and couple it with the attention mechanism. Guo et al (2018) utilize the interaction mechanism to weight the representations emitted by Bi-LSTM, and perform a deeper encoding by tensor network. Dai and Huang (2018) use Bi-LSTM to bring paragraph-level contextual information into argument representations. In addition, Qin et al (2016) build a hybrid neural model which couples two gated CNNs to extract both word-level and semantic-level convolutional features. Further, Qin et al (2017) integrate generative adversarial networks into multi-task learning network. Hereafter, Bai and Zhao (2018) establish multi-task network using multi-layer gated CNNs. The network is additionally coupled with residual networks and interactive attention mechanisms. Nguyen et al (2019) enhance Bai and Zhao (2018)'s multi-layer CNNs-based multi-task learning by minimizing the divergence between connective-level and relation-level embeddings.

5 Conclusion

The binary classification of implicit discourse relations still remains a great challenge. Although sophisticated representation learning models are deliberately used, the best performance achieved for a specific relation so far is less than 76% or even worse. We suggest that, in many cases, discourse relation recognition heavily relies on common-sense knowledge. In the future, we will attempt to acquire closely-related knowledge for attentive words and model their knowledge graphs. On the basis, attention-aware graph convolutional networks will be integrated into the existing IDRR architectures.

Acknowledgements

We are grateful for the comments of reviewers and thank all the co-authors for their considerable support. This work is not only supported by the national Key Research and Development Project of China via No. 2017YFB1002104 but national NSF of China via Grant No. 62076174, as well as the Stability Support Program of National Defense Key Laboratory of Science and Technology via Grant No. 61421100407.

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA*, pages 571–583.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Berlin, Germany*, pages 1726–1735.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA*, pages 141–151.

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, NAACL-HLT 2019, Minneapolis, MN, USA*, pages 4171–4186.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA*, pages 547–558.

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

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Douglas H Johnson. 1999. The insignificance of statistical significance testing. *The journal of wildlife management*, pages 763–772.

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, EMNLP 2017, Copenhagen, Denmark*, pages 1299–1308.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, Louisiana, USA*, pages 4848–4855.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*, pages 343–351.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France*.

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, EMNLP 2016, Austin, Texas, USA*, pages 1224–1233.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI 2016, Phoenix, Arizona, USA*, pages 2750–2756.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia*, pages 4068–4074.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *COLING*, pages 368–375.

Fandong Meng, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Interactive attention for neural machine translation. In *Proceedings of the 26th International Conference on Computational Linguistics, COLING 2016, Osaka, Japan*, pages 2174–2185.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar*, pages 1532–1543.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL 2009, Singapore*, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*.

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, EMNLP 2016, Austin, Texas, USA*, pages 2263–2270.

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

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45(11):2673–2681.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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, EMNLP 2015, Lisbon, Portugal*, pages 2230–2235.