Encoding and Fusing Semantic Connection and Linguistic Evidence for Implicit Discourse Relation Recognition

Wei Xiang[1], **Bang Wang**[1], **Lu Dai**[1], **Yijun Mo**[2*]
[1]School of Electronic Information and Communications,
Huazhong University of Science and Technology, Wuhan, China
[2]School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
{xiangwei, wangbang, dailu18, moyj}@hust.edu.cn

Abstract

Prior studies use one attention mechanism to improve contextual semantic representation learning for implicit discourse relation recognition (IDRR). However, diverse relation senses may benefit from different attention mechanisms. We also argue that some linguistic relation in between two words can be further exploited for IDRR. This paper proposes a *Multi-Attentive Neural Fusion* (MANF) model to encode and fuse both *semantic connection* and *linguistic evidence* for IDRR. In MANF, we design a *Dual Attention Network* (DAN) to learn and fuse two kinds of attentive representation for arguments as its semantic connection. We also propose an *Offset Matrix Network* (OMN) to encode the linguistic relations of word-pairs as linguistic evidence. Our MANF model achieves the state-of-the-art results on the PDTB 3.0 corpus.

1 Introduction

Implicit Discourse Relation Recognition (IDRR) is to detect and classify some latent relation in between a pair of text segments (called arguments) without an explicit connective word. It is of great importance for many downstream Natural Language Processing (NLP) applications, such as question answering (Liakata et al., 2013), machine translation (Guzmán et al., 2014), information extraction (Xiang and Wang, 2019), sentiment analysis (Wang and Wang, 2020), and etc. However, due to the absence of an explicit connective word, inferring discourse relations from the contextual semantics of arguments is still a challenging task.

Conventional machine learning based methods usually train a relation classifier by using many handmade features to capture lexical, syntactic regularity and contextual information of arguments, which is time-consuming and labor-intensive (Pitler et al., 2009, 2008). Deep learning based methods design diverse neural networks to automatic learn the contextual semantic representation of each argument, such as the Shallow Conventional Neural Network (SCNN) (Zhang et al., 2015), Tree-like Long Short-Term Memory (Tree-LSTM) (Rutherford et al., 2017), and BiLSTM-CNN framework (Guo et al., 2019). Although these neural networks can autonomously learn a kind of deeper contextual semantics of arguments, they do not differentiate arguments' words in the representation learning.

Recently, some attention mechanisms have been employed in neural networks to unequally treat words in representation learning. For example, the *self-attention* computes the local contextual importance of each word in one argument, which generally prioritizes *content words* for better learning substantive meaning of an argument (Zhou et al., 2016). The *interactive attention* weights each word in one argument according to its interaction with the representation of another argument, which usually focuses on the rhetorical device of two arguments, like prioritizing some *function words* with little substantive meaning but potentially indicating the connection of two arguments (Liu and Li, 2016; Guo et al., 2018).

Both kinds of attention mechanisms have been proven effective for IDRR, as each can well exploit either content semantics or rhetorical devices of an argument pair. We regard these contextual semantic information derived from argument content as a kind of *semantic connection* for relation recognition. However, the IDRR task normally needs to recognize diverse senses of relations, while different senses may benefit from different attentions. To enjoy both advantages, we propose to learn two kinds of argument representation, each based on one attention mechanism. They are next fused to encode an argument pair as the semantic connection for relation recognition in this paper.

Besides semantic connection, we argue that a

[*] Corresponding author: Yijun Mo

Findings of the Association for Computational Linguistics: ACL 2022, pages 3247 - 3257
May 22-27, 2022 ©2022 Association for Computational Linguistics

kind of *linguistic evidence* can be obtained from word distributed representation for relation recognition. Indeed, many pre-trained language models, like the word2vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019), are learned from a large amount of unlabeled text by encoding the linguistic regularities and patterns in an unsupervised way. As a word embedding contains inherent meaning of the word, it can be used to infer some linguistic relation in between two words by linear translation. This motivates us to encode such linguistic relations of word-pairs as linguistic evidence.

In this paper, we propose a *Multi-Attentive Neural Fusion* (MANF) model to encode and fuse both semantic connection and linguistic evidence for the IDRR task. The MANF model contains two modules. One is a *Dual Attention Network* (DAN) : It builds upon a BiLSTM to first encode a self-attentive representation and an inter-attentive representation for each argument. To adapt to different relation senses, we next use a fusion gate to integrate the two representations into the semantic connection representation. Another is an *Offset Matrix Network* (OMN): It first computes the offset between word embeddings of a word-pair that contains one word from the first argument and another word from the second argument. Upon the offset matrix, we next design an offset attention layer and a multilayer perceptron to encode the linguistic evidence representation. Finally, we design another fusion gate to integrate both semantic connection and linguistic evidence representation for relation recognition. Our MANF Model achieve the state-of-the-art results on the PDTB 3.0 corpus. Our main contributions are as follows:

- Propose a MANF model to encode and fuse semantic connection and linguistic evidence for the IDRR task.
- Propose a DAN to enjoy both self-attention and interactive attention for semantic connection encoding.
- Propose an OMN to encode word-pairs' offsets as linguistic evidence.
- Provide a new baseline result for the IDRR task on the PDTB 3.0 corpus.

2 The Multi-Attentive Neural Fusion Model

Fig. 1 illustrates our MANF model, including the DAN, the OMN, and a hierarchical fusion mechanism.

2.1 Dual Attention Network

Our DAN is built upon a BiLSTM or BERT to encode a self-attentive representation and an inter-attentive representation for each argument, which are next fused to output the semantic connection representation for an argument pair. We note that a BiLSTM has been widely used to capture word contextual semantics for its good sequential encoding capability. In our experiments, we also replace the word2ve by a fine-tuned BERT for comparison.

The DAN model is illustrated in the left part of Fig. 1, which consists of a BiLSTM layer, a dual attention layer, and a fusion gate layer. We use pre-trained word2vec word embeddings $\mathbf{x} \in \mathbb{R}^{d_w}$ to input the BiLSTM. An argument pair $(Arg_1; Arg_2)$ can be denoted by:

$$Arg_1 : [\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_{L_1}^1]; \tag{1}$$

$$Arg_2 : [\mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_{L_2}^2], \tag{2}$$

where $\mathbf{x}_i^1$ and $\mathbf{x}_j^2$ represents the i-th word embedding in the 1st argument and the j-th word embedding in the 2nd argument respectively, and d_w the word embedding dimension.

BiLSTM layer: After the BiLSTM, we obtain two hidden states $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ for each word in one argument from the forward and backward sequence respectively, which are concatenated to obtain an intermediate state $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i, \overleftarrow{\mathbf{h}}_i]$. We use a matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L]$ to denote an argument encoding after the BiLSTM, where $\mathbf{h}_i \in \mathbb{R}^{2d_h}$, $\mathbf{H} \in \mathbb{R}^{L \times 2d_h}$, d_h is the dimension of hidden state.

Dual attention layer: To enjoy both advantages of self-attention and interactive attention, we propose a dual attention mechanism to encode an argument pair. For self-attention, the representation of each argument $\mathbf{r_s}$ is formed by weighted sum of intermediate state vectors produced by BiLSTM (Zhou et al., 2016):

$$\boldsymbol{\alpha}_s = softmax(\mathbf{w}_s^\mathsf{T}\mathbf{H}), \tag{3}$$

$$\mathbf{r}_s = \mathbf{H}\boldsymbol{\alpha}_s^\mathsf{T}, \tag{4}$$

where $\boldsymbol{\alpha}_s \in \mathbb{R}^L$ is the self-attention weight vector of an argument computed by local contextual importance of each word, $\mathbf{w}_s$ a learnable parameter vector.

For interactive attention, we use the representation of one argument to weight each word in another argument (Ma et al., 2017; Meng et al., 2016). We sum up the intermediate states $\mathbf{h_i}$ to obtain an intermediate argument representation, i.e.,

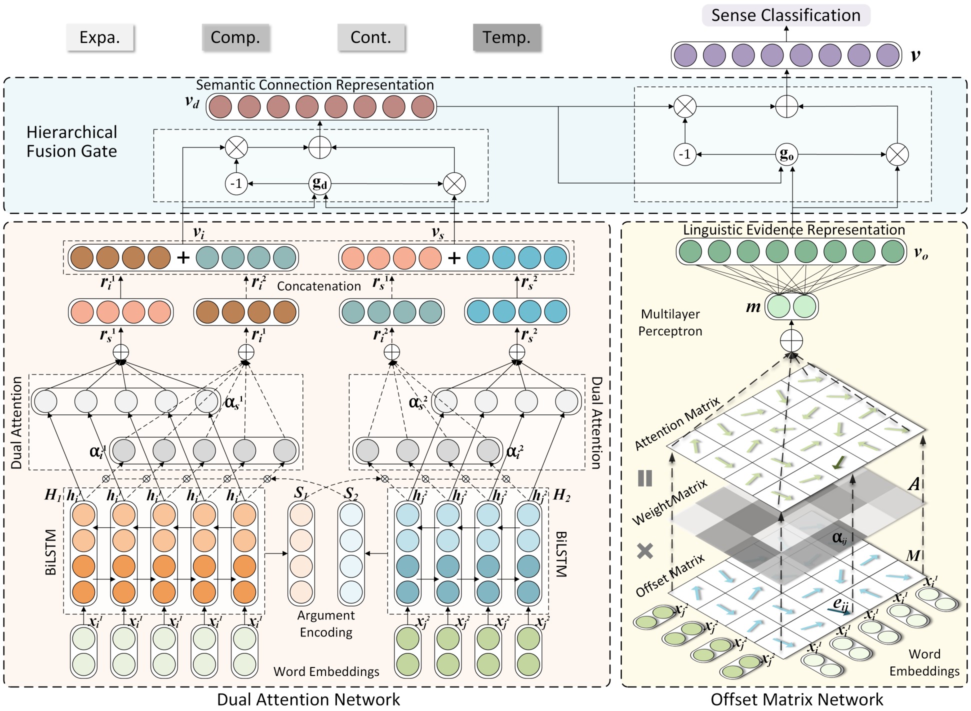


Figure 1: Illustration of our multi-attentive neural fusion model.

$\mathbf{S} = \sum_{i=1}^{L} \mathbf{h}_i$. The weight vector $\boldsymbol{\alpha}_i \in \mathbb{R}^L$ is computed by taking inner product between $\mathbf{S}$ and $\mathbf{H}$ cross two arguments, and followed by a softmax function as follows:

$$\boldsymbol{\alpha}_i^1 = softmax(\mathbf{H}_1 \mathbf{S}_2^\mathsf{T}) \tag{5}$$

$$\boldsymbol{\alpha}_i^2 = softmax(\mathbf{H}_2 \mathbf{S}_1^\mathsf{T}) \tag{6}$$

Finally, we weighted sum the intermediate state vectors with corresponding weight vector to form the interactive attention representation $\mathbf{r}_i$ for each argument:

$$\mathbf{r}_i^1 = \mathbf{H}_1(\boldsymbol{\alpha}_i^1)^\mathsf{T}, \quad \mathbf{r}_i^2 = \mathbf{H}_2(\boldsymbol{\alpha}_i^2)^\mathsf{T}. \tag{7}$$

Fusion gate layer: Considering the importance of the two attentions not always the same for different relation sense classification, we use a fusion gate to integrate their representations. First, we concatenate the representation of Arg_1 and Arg_2 to model their discourse relation as $\mathbf{v}_s = [\mathbf{r}_s^1, \mathbf{r}_s^2]$ and $\mathbf{v}_i = [\mathbf{r}_i^1, \mathbf{r}_i^2]$, where $\mathbf{v}_s, \mathbf{v}_i \in \mathbb{R}^{4d_h}$. The transition functions of fusion gate layer are computed as follows:

$$\mathbf{g}_d = sigmoid(\mathbf{W}_d \mathbf{v}_s + \mathbf{U}_d \mathbf{v}_i + \mathbf{b}_d), \tag{8}$$

$$\mathbf{v}_d = \mathbf{g}_d \odot \mathbf{v}_s + (1 - \mathbf{g}_d) \odot \mathbf{v}_i, \tag{9}$$

where $\mathbf{W}_d \in \mathbb{R}^{4d_h \times 4d_h}$, $\mathbf{U}_d \in \mathbb{R}^{4d_h \times 4d_h}$ and $\mathbf{b}_d \in \mathbb{R}^{4d_h}$ are learnable parameters and $\odot$ donates the element-wise product of vectors.

With the fusion gate, our DAN adaptively assigns different importance to self-attention and interactive attention, and outputs $\mathbf{v}_d \in \mathbb{R}^{4d_h}$ as the semantic connection vector for an argument pair.

2.2 Offset Matrix Network

We propose an OMN to encode the linguistic evidence representation based on the offsets of pretrained word embeddings, as shown in the right part of Fig. 1. First, we compute the offset between word embeddings of a word-pair that contains one word from the first argument and another word from the second argument. Then all the word-pair offsets of an argument pair compose an offset matrix $\mathbf{M} \in \mathbb{R}^{L_1 \times L_2 \times d_h}$, where $\mathbf{e}_{ij} \in \mathbb{R}^{d_h}$ is the offset vector between the i-th word in the 1st argument and the j-th word in the 2nd argument.

Considering that each word-pair in the offset matrix may have different contribution to the relation classification, we assign a weight score $\boldsymbol{\alpha}_{ij}$ to every offset vectors, and the weight scores are compute as follows:

$$\mathbf{A} = softmax(\mathbf{w}_o^\mathsf{T} \mathbf{M}), \tag{10}$$

where $\mathbf{A} \in \mathbb{R}^{L_1 \times L_2}$ is the weight matrix, $\mathbf{w}_o$ is a learnable parameter vector. We compute a word-pair interaction vector $\mathbf{m} \in \mathbb{R}^{d_h}$ as the weighted sum of all word-pair offset vectors:

$$\mathbf{m} = \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \mathbf{e}_{ij} \boldsymbol{\alpha}_{ij} \tag{11}$$

Next, we input $\mathbf{m}$ into a multilayer perceptron (MLP) followed by a tanh activation function to output the linguistic evidence vector $\mathbf{v}_o \in \mathbb{R}^{4d_h}$ for an argument pair:

$$\mathbf{v}_o = \tanh(\mathbf{W}_o \mathbf{m} + \mathbf{b}_o), \tag{12}$$

where $\mathbf{W}_o \in \mathbb{R}^{d_h \times 4d_h}$ and $\mathbf{b}_o \in \mathbb{R}^{4d_h}$ are learnable parameters.

2.3 Implicit Discourse Relation Classification

After obtaining the semantic connection vector $\mathbf{v}_d$ and the linguistic evidence vector $\mathbf{v}_o$, we also argue that they may have different importance for diverse relation sense classification. So we use another fusion gate to integrate the two kinds of representation vectors and obtain the final representation $\mathbf{v} \in \mathbb{R}^{4d_h}$ of an argument pair for relation classification. The transition functions are:

$$\mathbf{g}_o = sigmoid(\mathbf{W}_f \mathbf{v}_d + \mathbf{U}_f \mathbf{v}_o + \mathbf{b}_f), \tag{13}$$

$$\mathbf{v} = \mathbf{g}_o \odot \mathbf{v}_d + (1 - \mathbf{g}_o) \odot \mathbf{v}_o, \tag{14}$$

where $\mathbf{W}_f$, $\mathbf{U}_f \in \mathbb{R}^{4d_h \times 4d_h}$ and $\mathbf{b}_f \in \mathbb{R}^{4d_h}$ are learnable parameters.

The classifier is a fully connected layer with softmax to transform the final argument pair representation $\mathbf{v}$ to a probability distribution $\hat{\mathbf{y}} \in \mathbb{R}^n$ for predicting the discourse relation sense:

$$\hat{\mathbf{y}} = softmax(\mathbf{W}_c \mathbf{v} + \mathbf{b}_c),$$

where $\mathbf{W}_c \in \mathbb{R}^{4d_h \times n}$, $\mathbf{b}_c \in \mathbb{R}^n$ are learnable parameterss.

For model training, we adopt the cross entropy loss as the cost function:

$$J(\theta) = -\frac{1}{K} \sum_{k=1}^{K} \mathbf{y}^{(k)} \log(\hat{\mathbf{y}}^{(k)}) + \lambda \|\theta\|^2, \tag{15}$$

where $\mathbf{y}^{(k)}$ and $\hat{\mathbf{y}}^{(k)}$ are the gold label and predicted label of the k-th training instance respectively. λ and θ are the regularization hyper-parameters. We use the Adam optimizer and combine dropout with $L2$ regularization for model training.

Relation	Train	Dev.	Test
Expansion	8645	748	643
Comparison	1937	190	154
Contingency	5916	579	529
Temporal	1447	136	148
Total	17945	1653	1474

Table 1: Statistics of implicit discourse relation instances in PDTB 3.0 with four top-level relation senses.

3 Experiment Setting

3.1 The PDTB 3.0 Dataset

We conduct experiments on the latest version 3.0 of Penn Discourse TreeBank (PDTB) corpus, which was released on March 2019 and updated on February 2020. Following the conventional data splitting in PDTB 2.0, we use sections 2-20 as the training set, sections 21-22 as the testing set and 0-1 as the development set (Ji and Eisenstein, 2015). Our experiments are conducted on the four top-level classes of relation sense as the existing studies, including Comparison, Contingency, Expansion and Temporal. The statistics of implicit discourse relation instances in the PDTB 3.0 corpus are summarized in Table 1. More details about PDTB 3.0 are provided in the supplementary material.

3.2 Competitors

- **NNMA** (Liu and Li, 2016) combines two arguments' representation for stacked interactive attentions.
- **ANN** (Lan et al., 2017) applies interactive attention into a multi-task learning framework.
- **IPAL** (Ruan et al., 2020) propagates self-attention into interactive attention by a cross-coupled network.
- **DAGRN** (Chen et al., 2016b) encodes word-pair interactions by a neural tensor network.

3.3 Parameter Setting

We obtain the pre-trained word embeddings from the 300-dimensional English word2vec model ($d_w = 300$) provided by Google [1] and the 768-dimensional English BERT model ($d_w = 768$) provided by HuggingFace [2]. From our statistics, 99.46% of arguments do not exceed 50 words in PDTB3.0. So we set the maximum length of argument to 50 ($L = 50$). For the word2vec model, we set the mini-batch size to 32 and the

[1] code.google.com/archive/p/word2vec

[2] huggingface.co/bert-base-uncased

initial learning rate to 5e-4; while for the BERT model, the mini-batch size and initial learning rate is 16 and 1e-5. The trainable parameters are randomly initialized from normal distributions, and the dropout rate is set to 0.2 in the fusion gates and 0.5 in the MLP. We release the code at: https://github.com/HustMinsLab/MANF.

4 Result and Analysis

4.1 Overall Result

We implement four-way classification and binary classification (i.e. one-versus-others) on the PDTB 3.0, in which macro F_1 score and accuracy (Acc) are used for four-way classification and F_1 score is used for binary classification.

Table 2 compares the overall performance between our MANF and the competitors. In four-way classification, our MANF achieves significant improvements over competitors in terms of both macro F_1 and Acc. In binary-classification, ours also achieves the best performance in three relation sense classification, while the second with a small F_1 gap to that of NNMA in the Temporal sense classification.

We note that the first three competitors are neural models mainly for learning argument representation from contextual semantic connections; While the DAGRN is a neural model for learning representation from word linguistic evidences. The first observation is that the DARGN cannot outperform the first three competitors, though the performance gaps are not obvious. This might suggest that latent semantic connections learned from sequential contexts play the main role in relation recognition. This, however, is not unexpected. A relation is usually used for linking the meanings of two arguments, i.e., semantic connections, no matter with or without an explicit connective.

The second observation is that in the first three competitors, the ANN cannot outperform either NNMA or IPAL, not even once in all the performance metrics. We note that although they all employ attention mechanisms in learning semantic connection, the ANN applies a straightforward interactive attention to learn argument representation; While the NNMA designs a sophisticated mechanism for stacking multiple levels of attentions, and the IPAL employs a kind of sequential attention mechanism, i.e., interactive attention after self-attention.

Finally, we attribute the outstanding performance of our MANF model to its fusion of two attentions for learning semantic connection, as well as its exploitation of word linguistic evidence. This will be further analyzed in our ablation studies.

4.2 Ablation Study

Linguistic Evidence: We have argued that the inherent meaning of a word, other than its contextual semantics, can be exploited as a kind of linguistic evidence between two arguments for relation classification. To this end, we have designed the OMN module with the pre-trained word embeddings as its input. This input choice is from such considerations: A pre-trained word embedding is normally learned from a huge corpus containing materials from diverse backgrounds [3], which not only could capture some polysemous property for one word, but also could encode some linguistic regularity and pattern in between words from different contexts. While such properties might be compromised, if we input the OMN with the contextual semantic encodings.

To verify our arguments, we design two variants for the input of the OMN module. (1) **Shared**: It replaces the input of pre-trained $\mathbf{x}_i^1$ ($\mathbf{x}_j^2$) by the hidden state $\mathbf{h}_i^1$ ($\mathbf{h}_j^2$) of the respective BiLSTM in the DAN module. That is, two modules share the same BiLSTM for encoding word contextual semantics. (2) **Parallel**: We adopt additional BiLSM networks with their hidden states to replace pre-trained word embeddings. That is, two modules adopt parallel BiLSTM networks.

Table 3 presents the results of the three input choices for the OMN module. The better performance of using pre-trained word embedding can support our arguments. Although a BiLSTM network is well capable of encoding a word contextual semantics for its sequential processing mechanism, our design objective is to exploit the inherent meaning of a word to capture linguistic evidence for an argument pair. This is particular evident in the binary classification of Comparison and Temporal relation sense for its larger improvements. So using the pre-trained word embedding is a wise choice.

Module ablation study: To examine the effectiveness of different modules, we design the following ablation study.

[3] The word2vec was trained from the Google News dataset containing 100 billion words from diverse domain articles. The BERT was trained from the BookCorpus consisting of 11,038 books and English Wikipedia containing over six million articles.

Model	Four-way Classification		Binary Classification (F1)			
	F1	Acc	Expa.	Comp.	Cont.	Temp.
NNMA (EMNLP, 2016)	46.13%	57.67%	65.10%	29.15%	63.33%	**41.03%**
ANN (EMNLP, 2017)	47.29%	57.06%	64.03%	30.10%	60.91%	33.71%
IPAL (COLING, 2020)	49.45%	58.01%	64.28%	30.37%	61.95%	34.74%
DAGRN (ACL, 2016)	45.11%	57.33%	64.71%	27.34%	62.56%	38.91%
Our MANF	**53.14%**	**60.45%**	**67.82%**	**34.16%**	**65.48%**	40.22%

Table 2: Overall result of comparison models for implicit discourse relation classification.

Four-way Classification

Method	Pre-trained	Shared	Parallel
F1	**53.14** %	50.41%	51.54%
Acc	60.45%	58.82%	**60.85%**

Binary Classification (F1)

Method	Pre-trained	Shared	Parallel
Expa.	**67.82%**	67.13%	67.47%
Comp.	**34.16%**	31.43 %	30.48%
Cont.	**65.48%**	63.06%	64.93%
Temp.	**40.22%**	38.83%	38.36%

Table 3: Ablation study for linguistic evidence by using different word encodings as the OMN input.

- **BiLSTM (B)** is the building block of DAN, without two attentions and word-pair offsets.
- **B+SelfAtt** is a subpart of DAN, with only self-attention, but without interactive attention and word-pair offsets.
- **B+InterAtt** is a subpart of DAN, with only interactive attention, but without self-attention and word-pair offsets.
- **B+DualAtt (DAN)** is only the DAN module, with two attentions, but without word-pair offsets.
- **WordPair (OMN)** is only the OMN module, without argument representation for semantic connection.
- **B+WordPair** combines the OMN with a BiLSTM for encoding semantic connection but without any attention.
- **B+DualAtt+WordPair** is our MANF model.

Table 4 presents the results of our module ablation study. Among the first four models without using word-pair offsets, we first observe that the bare BiLSTM cannot outperform those employing attention(s) to differentiate words in argument representation learning. On the other hand, the B+DualAtt achieves better performance compared with the B+SelfAtt and B+InterAtt each using only one kind of attention, except a slight gap of Acc in the four-way classification. This indicates that our fusion of both attention mechanisms is an effective approach to augment semantic connection learning for an argument pair.

We also observe that the WordPair(OMN) only exploiting word-pair offsets performs the worst among all models. This, however, is not unexpected, as it totally ignores an argument semantics as well as latent semantic connection between arguments. On the other hand, the B+WordPair model, fusing linguistic evidence with semantic connection even learned by a bare BiLSTM without any attention, can greatly improve the performance of WordPair(OMN). The B+WordPair model can even achieve the best or the second best in some cases. This again validates our arguments of encoding and fusing both semantic connection and linguistic evidence to improve relation recognition.

Table 5 presents experiments using fine-tuned BERT to replace the word2vec based BiLSTM for semantic connection encoding. In contrast, the OMN module uses the BERT without fine-tuning to exploit linguistic evidence. The first three ablation modules correspond to the BiLSTM (B), B+DualAtt (DAN) and B+WordPair, respectively. We can observe that the BERT+DualAtt and BERT+WordPair models achieve better performance than the baseline BERT model. This further confirms the necessity of fusing both attention mechanisms and exploiting linguistic evidence. Finally, our MANF model yields substantial improvements overall ablation modules, and the outstanding performance approves our arguments and design objectives.

4.3 Case Study

We use case study to visualize and compare different attention mechanisms. Fig. 2 visualizes the word weight obtained by self-attention and interactive attention for four cases of different relation senses. We observe that the two attentions assign different weights to different words. In particular, the interactive attention seems to mainly focus on

Model	Four-way Classification		Binary Classification (F1)			
	F1	Acc	Expa.	Comp.	Cont.	Temp.
BiLSTM (B)	47.80%	57.67%	63.07%	28.05%	61.79%	36.40%
B+SelfAtt	49.39%	59.16%	66.79%	30.80%	64.72%	36.57%
B+InterAtt	50.70%	59.63%	67.30%	30.15%	62.36%	36.33%
B+DualAtt (DAN)	51.64%	59.50%	67.50%	32.18%	65.42%	38.53%
WordPair (OMN)	39.62%	51.22%	60.81%	25.95%	57.37%	26.87%
B+WordPair	50.95%	60.31%	67.01%	**34.30%**	63.15%	36.81%
B+DualAtt+WordPair (MANF)	**53.14%**	**60.45%**	**67.82%**	34.16%	**65.48%**	**40.22%**

Table 4: Experiment results of module ablation study.

Model	Four-way Classification		Binary Classification (F1)			
	F1	Acc	Expa.	Comp.	Cont.	Temp.
BERT	54.74%	62.69%	68.01%	34.75%	64.45%	40.25%
BERT+DualAtt (DAN)	55.23%	62.21%	68.18%	35.70%	65.07%	40.37%
BERT+WordPair	55.02%	61.67%	68.49%	**36.12%**	65.45%	**42.65%**
BERT+DualAtt+WordPair (MANF)	**56.63%**	**64.04%**	**70.00%**	35.83%	**66.77%**	42.13%

Table 5: Experiment results with the fine-tuned BERT language model.

one word with a very high weight in each argument, which is generally a kind of function word, such as the "*in*", "*back*" in the Temporal case, "*His*", "*and*" in Contingency case, and "*I*" in Comparison case. Such function words may be regarded as serving a kind of rhetorical devices for some common linguistic regularities and patterns.

On the other hand, the self-attention tends to assign several words in one argument with similar yet non-ignorable weights, which are often kinds of content words, such as "*slithered*", "*and*", "*slipped*" in the Expansion case. Such a few of content words might be more important to capture the contextual semantics of an argument, which can be next exploited for encoding semantic connection between two arguments. Such functionality differences of the two attentions indeed have motivated us to try a fusion mechaism, so as for each to excel in relation recognition of different senses.

Fig. 3 visualizes the weight matrix of word-pair offsets in the OMN module but with different input. It can be observed that using pre-trained word embeddings can help emphasizing the word-pair "*don't-did*" probably for their generally containing fewer contextual information. On the other hand, the other two using word contextual encoding pay attentions to word-pairs much similar to those words in the self-attention and interactive attention, such as "*I-I*", "*I-think*", "*I-don't*". As word contextual encoding has already been exploited in the DAN module, we argue that using pre-trained word embeddings for word-pair offsets could complete argument representation learning from another view of common linguistic evidence.

5 Related Work

The IDRR task is usually approached as a classification problem, and the key is the argument representation.

Machine learning approaches, like using a *Naive Bayes*, *Support Vector Machine* (SVM) classifier, have designed various features to capture lexical, syntactic regularity and contextual information as argument representation (Pitler et al., 2008; Lin et al., 2009; Pitler et al., 2009; Louis et al., 2010). However, manually crafting features is not only time-consuming and labor-intensive, but also suffers from data sparsity problem due to the use of one-hot feature encoding.

Deep learning models have prevailed for their capabilities of automatic learning argument representation (Zhang et al., 2015; Rutherford et al., 2017). For example, the SCNN model (Zhang et al., 2015) obtains each argument representation via a single convolution layer, and the concatenation of two arguments' representations is used for relation classification. Rutherford et al. (Rutherford et al., 2017) employ a LSTM network to capture word contextual semantics for argument representation. Some hybrid models have attempted to combine CNN, LSTM, graph convolutional networks and etc. for more sophisticated argument representa-

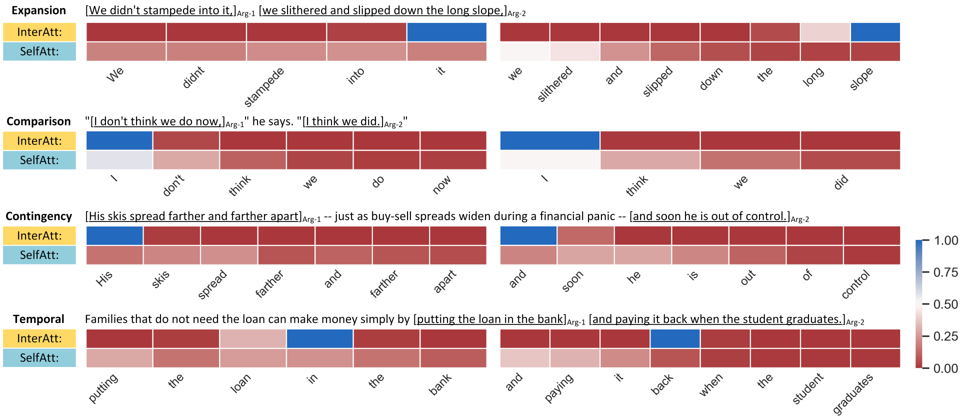
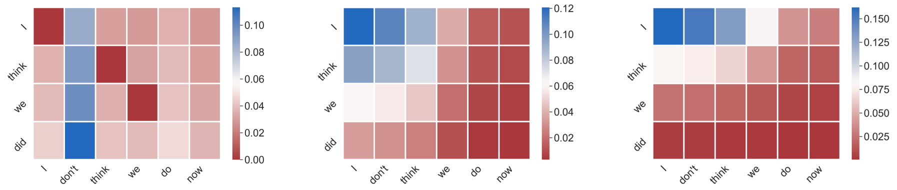


Figure 2: Visualization of attention weights for four cases of relations senses.



(a) Pre-trained word embedding (b) Shared BiLSTM (c) Parallel BiLSTM

Figure 3: Visualization of the weight matrix of word-pair offsets in the OMN module with the input of (a) pre-trained word embeddings, (b) hidden states of shared BiLSTM, and (c) hidden states of parallel BiLSTM.

tion (Guo et al., 2019; Xu et al., 2019; Zhang et al., 2021). These approaches, however, have ignored the fact that different words may contribute differently in argument representation learning.

Attention mechanisms can guide a neural model to unequally encode each word according to its contextual importance for argument representation (Zhou et al., 2016; Bai and Zhao, 2018; Liu and Li, 2016; Guo et al., 2018, 2020). For example, Zhou et al. (Zhou et al., 2016) apply self-attention to weight a word according to its similarity to its belonging argument. Guo et al. (Guo et al., 2018, 2020) adopt an interactive attention to differentiae words in one argument, where a word is weighted according to the similarity between its encoding and another argument representation. Liu and Li (Liu and Li, 2016) design a multi-level attention to repeatedly compute word importance in a hierarchical way. Ruan et al. (Ruan et al., 2020) propose a pipeline workflow to apply interactive attention after self-attention.

Word pair features have been exploited in machine learning and deep learning approaches for argument representation (Blair-Goldensohn et al., 2007; Biran and McKeown, 2013; Zhou et al., 2013; Chen et al., 2016a,b). For example, Biran and McKeown (Biran and McKeown, 2013) compute the appearance probabilities of aggregated word pairs to train a logistic regression classifier. Chen et al. (Chen et al., 2016b) construct a relevance score word-pair interaction matrix based on a bilinear model (Jenatton et al., 2012) and a single layer neural model (Collobert and Weston, 2008).

The proposed MANF model is a deep neural model, employing a hierarchical fusion mechanism to fuse two kinds of attentive word encodings as well as word pair offset encodings in argument representation learning.

6 Concluding Remarks

In this paper, we argue that implicit relation recognition can benefit from both semantic connection and linguistic evidence between arguments. Motivated from such considerations, we have designed the MANF model to encode and fuse them for the IDRR task. The MANF model consists a DAN module to fuse both self-attentive and inter-attentive contextual semantics for learning representation of semantic connection, and a OMN module to attentively encode word-pair offsets for learning representation of linguistic evidence. Both kinds of representations are finally fused for relation recognition. Experiments on the latest PDTB 3.0 corpus have validated our design objectives for the new benchmark performance established by our MANF model.

This paper has employed the pre-trained word embeddings trained by the word2vec and BERT; While other pre-training models shall also be adopted and tested in our future work. The performance differences of recognizing different relation senses also motivate to further investigate other advanced fusion mechanisms.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China (Grant No: 62172167). We also want to use our MANF model on MindSpore[4], which is a new deep learning computing framework. These problems are left for future work.

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*, pages 571–583, Stroudsburg, PA. The Association for Computational Linguistics.

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*, pages 69–73, Stroudsburg, PA. The Association for Computational Linguistics.

Sasha Blair-Goldensohn, Kathleen McKeown, and Owen Rambow. 2007. Building and Refining Rhetorical-Semantic Relation Models. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 428–435, Stroudsburg, PA. The Association for Computational Linguistics.

Jifan Chen, Qi Zhang, Pengfei Liu, and Xuanjing Huang. 2016a. Discourse relations detection via a mixed generative-discriminative framework. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2921–2927, Menlo Park, Calif. AAAI Press.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2016b. 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*, pages 1726–1735, Stroudsburg, PA. The Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, New York, NY. ACMPress.

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 Computational Linguistics*, pages 4171–4186, Stroudsburg, PA. The Association for Computational Linguistics.

Fengyu Guo, Ruifang He, and Jianwu Dang. 2019. Implicit discourse relation recognition via a bilstm-cnn architecture with dynamic chunk-based max pooling. *IEEE Access*, 7:169281–169292.

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. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7822–7829, Menlo Park, Calif. AAAI Press.

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*, pages 547–558, Stroudsburg, PA. The Association for Computational Linguistics.

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*, pages 687–698, Stroudsburg, PA. The Association for Computational Linguistics.

Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. 2012. A latent factor model for highly multi-relational data. In *Advances in neural information processing systems*, pages 3167–3175, Cambridge, MA. MIT Press.

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.

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, Stroudsburg, PA. The Association for Computational Linguistics.

[4] http://www.mindspore.cn/

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757, Stroudsburg, PA. The Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Stroudsburg, PA. The Association for Computer Linguistics.

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, Stroudsburg, PA. The Association for Computational Linguistics.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the SIGDIAL 2010 Conference*, pages 59–62, Stroudsburg, PA. The Association for Computer Linguistics.

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*, pages 4068–4074, Amsterdam, Netherlands. Elsevier.

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: Technical Papers*, pages 2174–2185, Stroudsburg, PA. The Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, Cambridge, MA. MIT Press.

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*, pages 683–691, Stroudsburg, PA. The Association for Computer Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K. Joshi. 2008. Easily Identifiable Discourse Relations. In *Proceedings of the 22th International Conference on Computational Linguistics*, pages 87–90, Stroudsburg, PA. The Association for Computer Linguistics.

Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. 2020. Interactively-Propagative Attention Learning for Implicit Discourse Relation Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3168–3178, Stroudsburg, PA. The Association for Computational Linguistics.

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*, pages 281–291, Stroudsburg, PA. The Association for Computational Linguistics.

Chang Wang and Bang Wang. 2020. An End-to-end Topic-Enhanced Self-Attention Network for Social Emotion Classification. In *Proceedings of The Web Conference 2020*, pages 2210–2219, New York, NY. ACMPress.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Sheng Xu, Peifeng Li, Fang Kong, Qiaoming Zhu, and Guodong Zhou. 2019. Topic Tensor Network for Implicit Discourse Relation Recognition in Chinese. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 608–618, Stroudsburg, PA. The Association for Computational Linguistics.

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

Yingxue Zhang, Fandong Meng, Li Peng, Jian Ping, and Jie Zhou. 2021. Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association Computational Linguistics*, pages 1592–1599, Stroudsburg, PA. The Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA. The Association for Computational Linguistics.

Xiaopei Zhou, Yu Hong, Tingting Che, Jianmin Yao, and Qiaoming Zhu. 2013. An unsupervised approach to inferring implicit discourse relation. *Journal of Chinese Information Processing*, 27(2):17–25.