EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Self-organizing incremental and graph convolution neural network for English implicit discourse relation recognition

Yubo Geng[1,*]

[1]School of English Language, Anhui International Studies University, Hefei, 230000 China

Abstract

Implicit discourse relation recognition is a sub-task of discourse relation recognition, which is challenging because it is difficult to learn the argument representation with rich semantic information and interactive information. To solve this problem, this paper proposes a self-organizing incremental and graph convolution neural network for English implicit discourse relation recognition. The method adopts the preliminary training language model BERT (Bidirectional Encoder Representation from Transformers) coding argument for argument. A classification model based on self-organizing incremental and graph convolutional neural network is constructed to obtain the argument representation which is helpful for English implicit discourse relation recognition. The experimental results show that the proposed method is superior to the benchmark model in terms of contingency relations and expansion relations.

Keywords: English discourse relation recognition, self-organizing incremental, graph convolution neural network, BERT.

Received on 12 November 2021, accepted on 20 November 2021, published on 22 November 2021

doi: 10.4108/eai.22-11-2021.172215

*Corresponding author. Email: byoungholee@qq.com

1. Introduction

Discourse relation recognition aims to study the logical relationship between two text segments (phrases, clauses, sentences or paragraphs) in the same discourse. As a basic research in the field of Natural Language processing (NLP), discourse relationship recognition is of great value in upper-level natural language processing applications [1,2], such as emotion analysis [3], machine reading comprehension [4], abstract extraction [5] and machine translation [6-8]. The task framework of discourse relation recognition is shown in figure 1. Given a pair of arguments (Arg1, Arg2), discourse relation classification model is used to identify discourse relations between them.

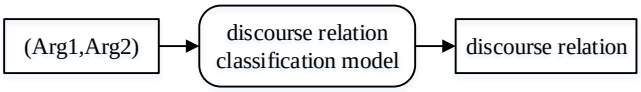

Figure 1. Task framework

At present, the largest authoritative corpus in the research field of discourse relationship recognition is Penn Discourse Tree-bank [9] (PDTB), which defines discourse relationship as a three-layer semantic relationship type system according to different granularity. The top four semantic relationships are comparison, contingency, expansion, and temporal. At the same time, according to whether there is a linking word (also known as a cue word, such as "because") between two argumentative expressions, PDTB divides discourse

relation into two categories: explicit discourse relation and implicit discourse relation. Explicit discourse relation is the type of discourse relation that can be inferred directly by explicit connectives. As shown in example 1, this explicit contingent relation argument pair contains the explicit conjunction "so, "a clue that Arg2 is the result of Arg1. Therefore, we can directly deduce that the argument pairs in example 1 have a contingent relation.

Example 1. [Arg1]: and will take measures
[Arg2]: **so** this kind of thing does not happen in the future
[Discourse relation]: Contingency. Cause. Result
In contrast, implicit discourse relation argument pairs lack explicit connectives, so they are more dependent on morphological, syntactic, semantic and contextual features. For example, the word "hurricane" in example 2 is the reason for the need for "precautionary mechanisms". Therefore, it can be inferred that the textual relationships in this thesis pair are fortuitous.
Example 2. [Arg1]: With a hurricane you know it is coming
[Arg2]: You have time to put precautionary mechanisms in place
[Discourse relation]: Contingency. Cause. Result

Explicit discourse relation studies have achieved high classification performance. Pitler et al. [10] had achieved 93.09% accuracy by using the mapping of explicit connectives and discourse relations. However, implicit discourse relation recognition performance is relatively low. The F1 values of the existing optimal methods in the four categories of relationships only reach 53%[11]. Therefore, this paper focuses on the task of implicit discourse relation recognition in English.

Previous studies have applied the attention mechanism to the calculation of argument representation [12-16] to evaluate the relevance of semantic information between arguments, so as to capture important semantic features to assist implicit discourse relationship recognition. However, relevant researches only focus on the semantic features of argumentative elements themselves or among them, so such a single feature cannot fully represent the semantic information of argumentative elements. If we focus only on the interaction information of argument, for example, the word pair information "good-wrong "and" ruined "in Example 3, it will easily lead to the argument pair being identified as a comparative relationship. But if an argument captures information about itself, looking at the words "not "and "good" in Arg1, and then looking at the word "ruined "in Arg2 with the interaction between arguments, then based on the words "not" and "ruined". The double negation of "(destroyed)" [17] can be inferred that the textual relation contained in this argument pair is contingent relation.
Example 3. [Arg1]: Psyllium's not a good crop
[Arg2]: You get a rain at the wrong time and the crop is ruined
[Discourse relation]: Contingency. Cause. Result

A graph convolutional neural network (SIG) based on self-organizing increment and interactive attention mechanism is proposed to construct implicit discourse relation classification model. This model constructs adjacency matrix based on self-organizing increment and inter-attention mechanism. Therefore, this model can utilize the semantic features of the argument itself and capture the interaction information of the argument, so as to encode a better representation of the argument and improve the performance of implicit discourse relation recognition.

In this paper, PDTB 2.0[2] data set is used for experiments and testing. The results show that the performance of the proposed model SIG is better than the benchmark model in English implicit discourse relation classification, and it is better than the current implicit discourse relation recognition model in many relations.

2. Related works

The existing researches on implicit discourse relation recognition mainly fall into two directions: constructing complex classification models and mining large amounts of training data. The model construction mainly includes machine learning model based on feature engineering and neural network model based on argument representation.
Previous studies have used a variety of linguistic features to construct statistical learning models.

On PDTB data set, Pitler et al. [18] attempted for the first time to use a variety of linguistic features to identify the top four implicit discourse relations, whose experimental performance exceeded that of random classification. Lin et al. [19] designed a discourse relation recognition model based on context features, word pair features, syntactic structure features and dependency structure features. Rutherford and Xue et al. [20] extracted Brownian clustering features to alleviate the sparsity of word pairs. Braud et al. [21] used the existing unsupervised word vector to train the maximum entropy model for implicit discourse relation classification based on shallow lexical features. Lei et al. [17] mined the semantic features of each relation, trained the naive Bayes model by combining the two cohesive means of topic continuity and argument source, and achieved a F1 value of 47.15% in four-way classification, whose performance exceeded most existing neural network models.

Most of the present researches on implicit discourse relation recognition build complex neural network models to improve the classification performance. Ji et al [22] used two recursive neural networks (RNN) to recognize implicit discourse relations based on the vector

representation of argument elements and entity fragments. Zhang et al. [23] proposed a shallow convolutional neural network containing only one hidden layer to avoid the over-fitting problem. Chen et al. [12] Based on bidirectional-long short-term memory network (Bi-LSTM), and semantic interaction information between word pairs was captured using gated relevancy network. Qin et al. [24] added gated neural network (GNN) on the basis of convolutional neural network to capture the interaction information (such as word pairs) between argument elements. Yin et al. [16] adopted the neural network model based on multi-task attention mechanism, and used the unlabeled external corpus BLLIP to generate pseudo-implicit discourse relation corpus to identify implicit discourse relation, and took it as an auxiliary task to improve PDTB implicit discourse relation recognition performance. Bai et al [13] constructed a complex argument representation model to extract argument features by integrating word vectors, convolution, recursion, residuals and attention mechanisms of different granularity. Nguyen et al. [11] used the model in reference [13]. In addition, based on knowledge transfer, relational representation and connective representation were mapped to make them in the same vector space, thus assisting implicit discourse relation recognition.

In view of the shortage of implicit discourse relation corpus, different methods have been used to expand the implicit corpus of PDTB. Zhu et al. [25] mined instances consistent with the original corpus in terms of semantics and relations from other data resources through argument vector. Wu et al. [26] found that explicit and implicit mismatch exists in bilingual corpus, that is, there were no connectives in English corpus, but there were explicit connectives in corresponding Chinese corpus. Based on this, Wu et al. [26] extracted pseudo-implicit discourse relation corpus from the CORPUS of FBIS and Hong Kong Law. Xu et al. [27] used explicit discourse relational corpus to construct pseudo-implicit style examples, and selected samples with high information content based on active learning method to expand implicit discourse relational corpus. Ruan et al. [28] used the "WHY" question pair in question answering corpus to generate pseudo-implicit argument pairs based on "declarative conversion of questions" so as to expand implicit causality corpus.

3. Methodology

The graph convolutional neural network (SIG) framework based on self-organizing increments and interactive attention proposed in this paper is shown in figure 2.

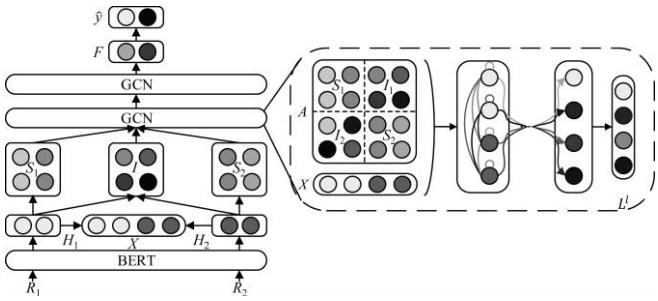

Figure 2. SIG model

Firstly, the argument representation of two arguments is obtained by fine-tuning BERT language model [29]. Secondly, a fully connected word-word graph is obtained by spliced feature matrix and adjacency matrix. As the initial features of graph convolutional network (GCN), word features are convolved and nonlinear transformation operations are performed on the hidden layer of double-layer GCN to obtain the final word representation. Finally, the word representation is sent to the full connection layer for dimensionality reduction, and the softmax function is used to normalize it, and the final classification result is obtained.

3.1. Vector representation layer

Given a deterministic metarepresentation $R_1 = (x_1^1, x_2^1, \cdots, x_L^1)$ and $R_2 = (x_1^2, x_2^2, \cdots, x_L^2)$, this paper uses fine-tuned BERT pre-trained language model to encode it. Specifically, R_1 and R_2 are spliced as model inputs to obtain the argument distributed representation $H = (h_1, h_2, \cdots, h_{2L+3})$. Where, $h_i \in R^{dk}$ represents the vector representation of the i-th word encoded by BERT after splicing. Finally, according to the maximum length of argument L, the encoded argument is extracted from H to presentation H_1 and H_2, and the specific calculation is shown in equations (1)~ (3).

$$H = BERT([CLS, R_1, SEP, R_2, SEP]) \quad (1)$$

$$H_1 = (h_2, h_3, \cdots, h_{L+1}) \quad (2)$$

$$H_2 = (h_{L+3}, h_{L+4}, \cdots, h_{2L+2}) \quad (3)$$

CLS is a special classification symbol, and its BERT encoded vector representation can be used as the vector representation of the whole input sequence. SEP is a special symbol used to separate two arguments in an input sequence.

3.2. Self-organizing Incremental Graph Convolutional Neural Network (WSOINNGCN)

The WSOINNGCN model framework is shown in figure 3, which consists of three parts: The first part is the feature vector set of image data obtained based on transfer learning; In the second part, a self-organizing incremental neural network (WSOINN) is used to extract topology structure of feature data [31,32], and a few nodes are selected for manual annotation according to the number of node victories. In the third part, graph convolution network (GCN) is built. The cross entropy loss function and Adam algorithm are used to optimize the network parameters, and the remaining nodes are automatically labeled. Finally, all image data are classified based on Euclidean distance.

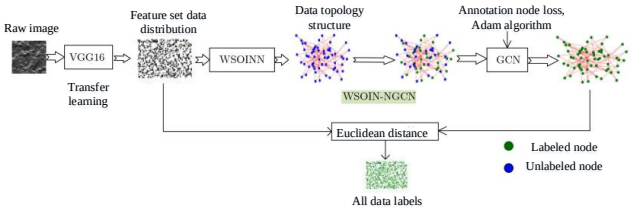

Figure 3. WSOINNGCN model

As shown in figure 4, the VGG16 convolutional module trained on ImageNet data set is used to extract the features of each text, and the 512 feature graphs obtained are pooled globally by means of means. Each graph outputs a 512-dimension feature vector, so as to obtain the data feature set after extraction of all text features.

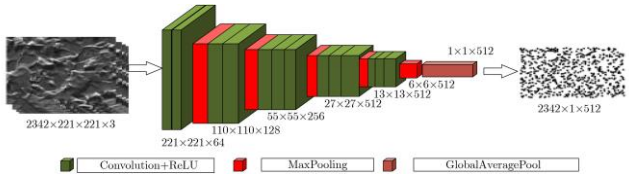

Figure 4. VGG16 convolution module is used to extract text features

A self-organizing incremental neural network with connection weight policy is introduced (WSOINN)

SOINN can obtain the spatial topological graph structure of feature data, while GCN can be used to mine the relationship of huge, sparse and super-dimensional association graph data. In order to integrate SOINN and GCN, this paper proposes the introduction of self-organizing incremental neural network (WSOINN) with connection weight number, and the introduction of node victory times to select a small number of nodes for manual annotation. The algorithm steps of WSOINN are as follows:

(1) Initialize the node set $V=\{v_1,v_2\}$, $v_1,v_2\in R^d$. Connection $E\subseteq N\times N$ is empty set. Node win number Win_times ={ t_{v_1} =0, t_{v_2} = 0}.

(2) Receive the new input sample $\xi\in R^d$, search the nodes s_1 and s_2 closest to ξ in V according to the Euclidean norm, i.e. $s_1=\arg\min_{v_n\in V}\|\xi-v_n\|$, $s_2=\arg\min_{v_n\in V\setminus\{s_1\}}\|\xi-v_n\|$, the win number is added 1, i.e. $t_{s_1}=t_{s_1}+1$, $t_{s_2}=t_{s_2}+1$.

(3) Calculate the similarity threshold T_{s_1}, T_{s_2} of nodes s_1 and s_2. For node $v\in V$, the set of nodes connected with v is denoted as P, if $P=\phi$, $T_v=\min\|v-v_n\|$, $v_n\in V\setminus\{v\}$, if $P\neq\phi$, $T_v=\max\|v-v_n\|$, $v_n\in P$.

According to the above WSOINN algorithm process, the connection weight W between the nodes represents the similarity between the two nodes, and the larger the connection weight is, the more similar the two nodes are. The more victories a node has, the more representative and important it is.

Nodes features matrices

Given two encoded argument representations H_1 and H_2, they are spliced as node characteristic matrix $X\in R^{2L\times dk}$, i.e., $X=[H_1,H_2]$. On this basis, the graph convolution operation can be performed on the two argument representations at the same time, so as to obtain the characteristic matrix rich in the argument's own information and interactive information.

Adjacency matrix

Considering that textual relations depend on deep text understanding and information interaction between arguments, this paper constructs the adjacency matrix of graph convolution neural network based on the self attention score matrix and interactive attention score matrix of arguments, so as to obtain a fully connected graph with arguments as nodes. The calculation methods of self attention mechanism and interactive attention mechanism used in this paper are introduced below.

In this paper, the self attention mechanism [32] is used for argument representations H_1 and H_2 to measure the importance of each word representation, so as to obtain the self attention score matrix $S \in R^{L\times L}$ of argument. Taking Arg1 as an example, the specific calculation is shown in equations (4) to (6).

$$Q_1 = H_1 W_{Q1} \tag{4}$$

$$K_1 = H_1 W_{K1} \tag{5}$$

$$S_1 = soft\max(\frac{Q_1 K_1^T}{\sqrt{d_k}}) \tag{6}$$

Where $W_{Q1} \in R^{d_k \times d_k}$ and $W_{K1} \in R^{d_k \times d_k}$ are learnable parameter matrixes. We take $\sqrt{d_k}$ as the denominator to prevent the inner product from being too large. Similarly, the self attention weight distribution matrix S_2 of Arg2 can be calculated.

At the same time, after the vector representation H_1 and H_2 of the two argument pairs are obtained, the interactive attention mechanism is used to calculate the interactive attention matrix $I \in R^{L\times L}$ of the argument pairs. Specifically, the normalization of I can obtain the interactive attention score a of Arg1 to Arg2 for each word in Arg2. Similarly, the normalization of I can obtain the interactive attention score I2 of Arg2 to Arg1 for each word. The specific calculation is shown in equations (7)~ (9).

$$I = H_1 W_1 H_2^T \tag{7}$$

$$I_1 = soft\max(I) \tag{8}$$

$$I_2 = soft\max(I^T) \tag{9}$$

Where, the learnable parameter matrix $W_I \in R^{d_k \times d_k}$ is the medium of Arg1 and Arg2 information interaction.

Through the above calculation, the self attention score matrices S_1 and S_2 and the interactive attention score matrices I_1 and I_2 can be obtained. Based on this, this paper splices S_1, S_2, I_1 and I_2 to obtain the adjacency matrix $A \in R^{2L\times 2L}$ integrating the argument's own information and interactive information, the specific splicing method is shown in equation (10).

$$A = \begin{pmatrix} S_1 & I_1 \\ I_2 & S_2 \end{pmatrix} \tag{10}$$

Graph convolution operation

The node feature matrix and adjacency matrix A of graph convolutional neural network are obtained based on the above formula. We refer to formula (4) to calculate the graph convolution feature of node feature matrix X. The number of GCN layers is 2, and the specific calculation is shown in formula (11).

$$L^2 = f(Af(AXW_1 + b_1)W_2 + b_2) \tag{11}$$

3.3. Full connection layer

In this paper, the updated feature representation $L^l = \{g_1^l, g_2^l, \cdots, g_{2L}^l\}$ is obtained through multi-layer GCN, where $g_i^l \in R^{d_k}$ represents the feature representation of the i-th node updated by layer l-th GCN. In this paper, the feature representation of each node output by the GCN of the last layer is summed to obtain the final argument pair feature representation $F \in R^{d_k}$. The specific calculation is shown in equation (12).

$$F = \sum_{i}^{2L} g_i^l \tag{12}$$

By inputting F into the full connection layer, we calculate the probability of relation r between Arg1 and Arg2, as shown in formula (13).

$$\hat{y} = soft\max(WF^T + b) \tag{13}$$

Where $W \in R^{n\times d_k}$, $b \in R^n$ are learnable parameters, W can reduce the dimension of the final feature representation F. $\hat{y} \in R^n$ is the probability of predicting whether this argument pair has a relation r.

3.4. Training

This paper constructs a binary classifier for each of the four class relationships of PDTB corpus. In the training process, this paper uses the cross entropy loss function as the objective function and uses Adam [33-35] optimization algorithm to update all model parameters. For a given argument pair (Arg1, Arg2) and its relationship label y_i, the loss function is calculated as shown in equation (14).

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{14}$$

Where $\hat{y}_i$ refers to the probability of whether there is a relation r between pairs of arguments. Since softmax activation function is used in this paper, $\hat{y}_i > 0$, $\sum_{i=1}^{n} \hat{y}_i = 1$. $y_i \in [0,1]$ indicates whether the argument pair has a true label. n indicates the number of categories.

4. Experiments

4.1. Experimental data

In this paper, the experiment of implicit text relation recognition is carried out with SIG model on the corpus of Pennsylvania text tree bank (PDTB). PDTB was proposed by Prasad in 2008, it came from 2304 articles in the Wall Street Journal (WSJ), and a total of 40600 text relationship samples were marked, 16224 samples were implicit text relationship examples. In order to keep consistent with the previous work, this paper takes section 02-20 as the training set, section 00-01 as the development set and section 21-22 as the test set. The data distribution of the top four semantic relationships comparison (COM.), continuity (CON.), expansion (EXP.) and temporary (TEM.) are shown in Table 1.

Table 1. PDTB data distribution of four kinds of implicit discourse relation

Relation type	Training set	Development set	Testing set
COM.	1855	189	145
CON.	3240	280	275
EXP.	6675	640	530
TEM.	580	50	55
Total	12350	1159	1005

It can be seen from table 1 that in the PDTB data set, the amount of text relationship data of the other three categories except EXP. is small, and the problem of inter class imbalance makes researchers usually train two classifiers separately for each relationship type for evaluation. Therefore, referring to the previous work, this paper trains the binary classification model based on the training sets of different text relations, and obtains a total of four binary classifiers, which are respectively used to judge whether the sample contains the text relation, and evaluates its performance through F1 value. Following the previous work, this paper does not integrate the four secondary classification results of the same sample, and only discusses the yes or no problem of single category text relationship in the secondary classification. In addition, because the PDTB data set has the problem of unbalanced positive and negative samples, this paper randomly down samples the negative samples to construct a training data set with balanced positive and negative samples. At the same time, in order to better compare with previous work, this paper carries out four-way classification experiments on PDTB data set, trains a four classifier based on the training set, and evaluates it with Macro-F1 value and accuracy.

4.2. Experimental setting

In order to prove that using GCN to fuse self attention and interactive attention mechanism is helpful to implicit text relationship recognition, the following six comparison systems are set up in this paper.

1) Bert (baseline): after the hidden layer outputs of Arg1 and Arg2 are obtained by fine tuning the Bert model, they are cut respectively to obtain the representation of two arguments. Then the sentence level argument representation is obtained by word by word summation, and the final feature is obtained by splicing the two sentence level representations. And it is input to the full connection layer for classification.

2) Self: after using Bert to obtain the argument representation of Arg1 and Arg2, calculate their self attention scores respectively, and apply the self attention weight to the argument representation; Then, the updated argument representation is summed word by word to obtain sentence level representation; Finally, the sentence level representation is spliced as the input of the whole connection layer.

3) Inter: after the argument representation of Bert output is obtained, the interactive attention mechanism is used to obtain the interactive attention weight distribution matrix and act on the argument representation; Then, the sentence pair level argument representation is obtained by summing and splicing the new argument representation word by word, and the full connection layer is input for implicit text relationship classification.

4) Concatenate: the sentence level argument representation is obtained by splicing the sentence level representations generated by the above self and inter systems, and input into the full connection layer for implicit text relationship classification.

5) Transformer: splice the argument representations of Arg1 and Arg2 obtained through Bert coding as the input of the double-layer transformer [36,37] with eight-head attention mechanism, and then sum the word features encoded by transformer word by word to obtain the sentence level representation of argument pairs, and input them to the full connection layer for implicit text relationship classification.

6) SIG: after the argument representations of Arg1 and Arg2 are obtained by Bert, the self attention weight distribution matrix and interactive attention weight distribution matrix are calculated respectively. Then, the two argument representations are spliced to obtain the

characteristic matrix, and then the attention weight distribution matrix is spliced to obtain the adjacency matrix to construct the double-layer GCN. The output of the last GCN layer is summed word by word to obtain the sentence level representation of two arguments, which are input into the full connection layer for implicit text relationship classification.

4.3. Parameter setting

In this paper, the output of the hidden layer of the fine tuned Bert is used as the argument representation, where we set the hidden layer vector dimension d_k to 768 and the maximum argument length L to 80. Based on the characteristic matrix constructed by argument representation, this paper splices argument self attention and interactive attention weight distribution matrix to obtain adjacency matrix, constructs two-layer (L=2) GCN neural network, and uses tanh function as the activation function of the model. When building the transformer model, we use the encoder of Transformer in the work of Subakan et al. [32] as a layer of Transformer in this paper. In this paper, a two-layer Transformer is used to transform the argument representation after coding, and the hidden layer dimension of the feed-forward neural network is set to 768, and GeLU [38] is used as the activation function. In the training process, the cross entropy is used as the loss function, and the batch gradient descent method based on Adam is used to optimize the model parameters, in which the batch size is 32 and the learning rate is $5e^{-5}$. In this paper, dropout is calculated after the last GCN layer, and the probability of random discarding is 0.1.

4.4. Experimental results

Six neural network models with different structures are used to classify the four categories of implicit text relations of PDTB. The specific classification performance is shown in Table 2. Among them, the performance of the proposed model sig in multiple relationships is better than the other five comparison models. The main reason is that sig combines the advantages of two attention mechanisms. While paying attention to the information of two arguments, it can also pay attention to the interactive information between them, and update the argument representation through such information. Therefore, SIG can generate argument representation that is more consistent with the characteristics of implicit text relationship classification task.

Table 2. Classification results of four categories of discourse relations by different models/%

Model	COM.	CON.	EXP.	TEM.
BERT	41.25	55.78	73.50	35.45
Self	41.21	57.17	73.51	39.05
Inter	43.86	56.31	73.58	39.31
concatenate	41.64	53.67	73.29	37.35
Transformer	46.94	56.87	74.73	41.71
SIG	48.19	60.81	74.60	42.11

However, the model transformer uses an 8-head attention mechanism to capture various information of arguments themselves and the interaction between arguments. However, when transformer simulates the information interaction between arguments, it only uses the argument point product matrix as the attention score matrix, while the attention mechanism that SIG can use is more flexible. In this paper, bilinear model is used to simulate the linear interaction between two arguments. In addition, transformer uses 8-head attention mechanism, while SIG only uses single head self attention mechanism; At the same time, the value of transformer's attention score matrix is inconsistent in different layers, while GCNs in different layers in SIG share the same adjacency matrix, and the size of its element value indicates the strength of the connection between different word nodes; After each layer transformer uses the attention mechanism to update the argument characteristics, it also needs to use the feed-forward neural network containing two fully connected layers to transform it, and adopts the residual mechanism. In contrast, the structure of SIG model is simpler and prevents over fitting to a certain extent. Therefore, transformer performs better than SIG in the expansion relationship with a large amount of data, but slightly weaker in other relationships.

In addition, the performance of concatenate model is inferior to self and Inter in almost all discourse relations. We believe that it is mainly caused by the following two reasons: first, the way of splicing is too simple to simulate the complex relationship between the two arguments and the balance between the two attention mechanisms; Secondly, there is a certain over fitting problem in this model. In contrast, the proposed model SIG uses GCN to weigh the two attention mechanisms. Among them, the inherent weight sharing characteristics of GCN model can prevent over fitting to a certain extent, so SIG can almost surpass other models in the classification performance of four types of text relationships.

In order to prove the effectiveness of the model SIG proposed in this paper, we compared it with the existing advanced models (see Table 3). Among them, Bai et al [13] used character level, sub word level and word level representation based on Shahid [35] to construct multi-granularity argument representation, and combined convolution operation, residual mechanism, interactive attention mechanism and multitask learning idea to construct complex deep neural network. On the basis of Bai et al [13], Nguyen et al. [11] mapped the relationship vector and conjunction vector to the same vector space based on the idea of knowledge transfer. In addition, Yin et al. [16] trained the multitasking model with the help of external data such as BLLIP. In the same text, there is a certain relationship between top-down text relationships. Xu et al [36] deeply explored this feature and constructed an implicit text relationship classifier by using the method of ensemble learning.

Table 3. Comparison results of SIG and existing advanced models/%

Model	Binary				Four-way classification	
	COM.	CON.	EXP.	TEM.	Macro-F1	Accuracy
Zhang [23]	33.23	52.05	69.61	30.55	---	---
Chen [12]	40.18	54.77	53.21	31.33	---	---
Qin [24]	41.56	57.33	71.51	35.44	---	---
Liu [15]	37.92	55.89	69.98	37.18	44.99	57.28
Yin [16]	40.74	58.97	72.48	38.51	47.81	57.41
Xu [36]	46.81	57.11	70.42	45.62	48.83	57.45
Lei [17]	43.25	57.83	72.89	29.11	47.16	---
Shi [14]	40.36	56.82	72.12	38.66	47.59	59.07
Bai [13]	47.86	54.48	70.61	36.98	51.07	---
Nguyen [11]	48.45	56.85	73.67	38.61	53.01	---
SIG	48.09	60.71	74.51	42.01	52.52	60.19

Compared with previous work, the model SIG proposed in this paper is relatively simple, and only standard PDTB data set is used for training. However, it can outperform the current optimal method in classification performance on multiple relations. The main reasons are as follows: 1) BERT pre-trained language model [39-42] already contains a large amount of prior knowledge, which is helpful for implicit discourse relation recognition which requires common knowledge. 2) Previous work usually uses interactive attention mechanism to extract interaction information between elements, but ignores the importance of the information of the elements themselves, while SIG integrates its own information and interaction information.

Table 4 shows the lexical distribution of the four categories of PDTB used in this paper. Where, each type of relationship contains a large number of out-of-vocabulary (OOV). Researchers usually represent these unregistered words with the special symbol "UNK" and uniformly initialize them to obtain a consistent word vector, which can break the dilemma of finding unregistered word vectors, but it reduces a certain amount of information and has a certain impact on implicit text relationship recognition.

Table 4. Lexical distribution in four categories of discourse relations

Data set	COM.	CON.	EXP.	TEM.
Training set	17005	22518	30612	10694
Testing set	6616	6616	6616	6616
Unregistered word	2010	1668	1276	2837

For example, in example 4, the unregistered word "steamed "does not appear in the training set. In the absence of the word "paused" and "reaching its high ", it is difficult to derive the causal relationship. However, BERT can use the word context information to initialize the word vector for the unregistered word, and "steamed Forward "is the reason for "reaching its high". Therefore, it can be deduced that the discourse relation contained in this argument pair is contingent relation.

Example 4 [Arg1]: Instead, the rally only paused for about 25 minutes and then **steamed** forward as institutions resumed buying.

[Arg2]: The market closed minutes after reaching its high for the day of

[Discourse relation]: Contingency. Cause. Result

In order to prove the effectiveness of model SIG, this paper uses models Self, Inter and SIG to calculate the distribution of attention weight for example 3, and average the value of attention weight word by word to draw gray color blocks, and obtain the grayscale of attention distribution calculated by example 3 for the three models (See figure 5). As can see from figure 5, both model Self and SIG focus on the words "not "and "good" in Arg1. However, only model SIG gives a high weight to the word "ruined "in Arg2. Thus, model SIG can infer from the double negation of the word "not "and "ruined" that the implicit discourse relationship contained between these two arguments is contingent.

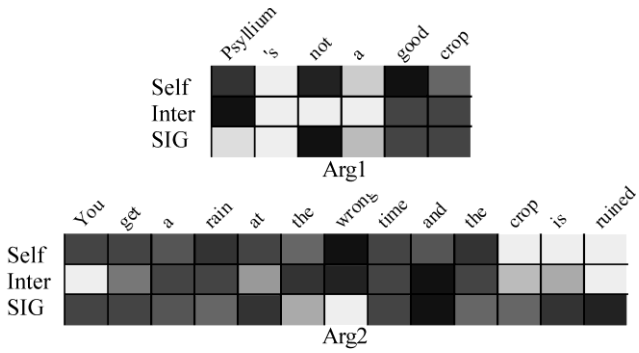

Figure 5. Example 3 gray scale of attention distribution obtained from different systems

In this paper, experiments are carried out on the model constructed by GCN with different layers, and its performance is shown in Table 5.

Table 5. Model classification performance based on GCN at different layers/%

Layer	Binary				Four-way classification	
	COM.	CON.	EXP.	TEM.	Macro-F1	Accuracy
GCN1	44.28	57.99	74.19	41.71	50.39	57.50
GCN2	48.09	60.71	74.50	42.01	52.52	60.19
GCN3	47.31	58.70	74.11	41.75	53.48	61.29
GCN4	47.44	56.91	74.15	41.71	53.87	59.49
GCN5	44.31	57.19	73.85	39.83	52.46	59.29
GCN6	44.61	56.48	73.59	38.43	51.44	59.69

Where, when the number of GCN layers is 2 (i.e. GCN2), the binary classifier reaches the maximum value in F1 value, while when the number of GCN layers is 4, the macro-F1 value and the accuracy of four-way classification are 53.87% and 59.49%, respectively. This is mainly because the sample size of the training set of the binary classification model is lower than that of the four-classification model. Therefore, when the number of GCN layers is large, the binary classifier tends to over-fit.

5. Conclusions

In this paper, a graph convolutional neural network model based on self-organizing increment and interactive attention mechanism is proposed to recognize implicit discourse relations. Experimental results show that the performance of the proposed model SIG is better than that of the benchmark model BERT, and the performance of the proposed model SIG is better than that of the existing advanced methods on multi-class relationships. The

experimental results show that the implicit discourse relation recognition task is still very challenging, and the classification performance of the other three categories except EXP. is low, which is far from meeting the requirements of practical application. In the next step, we will carry out researches from two aspects: (1) mine high-quality implicit discourse relation corpus externally for data imbalance. (2) construct a more complex classification model conforming to the characteristics of implicit discourse relation recognition tasks.

Acknowledgements.

This work was funded by “Key project of Humanities and Social Science research in Universities of Anhui Province in 2020: From the perspective of Translator's Behavior criticism & LT; Water margin & GT; A Contrastive Study on the Translation of Cultural Carrier words into English: A case study of pearl S. Buck and Charles Shapiro. Project Number: SK2020A0638. Education Department of Anhui Province”. Also supported by “2020 Provincial Quality Project of Universities in Anhui Province: Application of intelligent Classroom in English Listening and speaking Teaching based on "super star learning pass", Project Number: 2020jyxm0728, in Education Department of Anhui Province”.

References

[1] R. Geng, P. Jian, Y. Zhang and H. Huang, "Implicit discourse relation identification based on tree structure neural network," 2017 International Conference on Asian Language Processing (IALP), 2017, pp. 334-337, doi: 10.1109/IALP.2017.8300611.

[2] Huibin Ruan, Yu Sun, Yu Hong, et al. Graph Convolutional Network Based Implicit Discourse Relation Recognition [J]. Journal of Chinese Information Processing. 35(8), 28-37, 2021. (In Chinese)

[3] Zhou L, Li B, Gao W, et al. Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities[C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 162-171, 2011.

[4] Baynes K, Davis C H, Long D L. Comprehension of discourse relations in the right and left cerebral hemispheres[J]. Brain & Language, 2005, 95(1):111-112.

[5] Yoshida Y, Suzuki J, Hirao T, et al. Dependencybased discourse parser for single-document summariza. tion[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 18341839.

[6] Joty S, Guzman F, Marquez L, et al. DiscoTK: Using Discourse Structure for Machine Translation Evaluation[J]. 2019. arXiv:1911.12547

[7] Shoulin Yin, Hang Li, Asif Ali Laghari, et al. A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection[J]. EAI Endorsed Transactions on Scalable Information Systems. 21(33), e8, 2021. http://dx.doi.org/10.4108/eai.6-10-2021.171247

[8] Dongling Wang, Xiaowei Wang, and Shoulin Yin. A New Recursive Neural Network and Center Loss for Expression Recognition [J]. International Journal of Electronics and Information Engineering. Vol. 13, No. 3, pp. 97-104, 2021.

[9] Prasad R, Dinesh N, LeeA, et al. The Penn Discourse TreeBank 2.0 [C]//Proceedings of the International. Conference on Language Resources and Evaluation, 2008: 2961-2968,

[10] Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., Joshi, A.: Easily Identifiable Discourse Relations. In: Proceedings of the 22nd International Conference of Computational Linguistics (COLING 2008), Manchester, UK, pp. 87–90 (2008)

[11] Linh The Nguyen, Linh Van Ngo, Khoat Than,et al. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings[C]//Proceedings of the 57 th Annual Meeting of the Association for Computational Linguistics, 2019: 4201-4207.

[12] Chen J, Zhang Q, Liu P, et al. Implicit discourse relation detection via a deep architecture with gated relevance network[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, 1: 1726-1735.

[13] Bai H, Zhao H. Deep enhanced representation for implicit discourse relation recognition[J]. Xiv:1807.05154, 2018.

[14] Qingwu Shi, Shoulin Yin, Kun Wang, et al. Multichannel convolutional neural network-based fuzzy active contour model for medical image segmentation. Evolving Systems (2021). https://doi.org/10.1007/s12530-021-09392-3

[15] Desheng Liu, Linna Shan, Lei Wang, Shoulin Yin, et al. P3OI-MELSH: Privacy Protection Point of Interest Recommendation Algorithm Based on Multi-exploring Locality Sensitive Hashing[J]. Frontiers in Neurorobotics, 2021. doi: 10.3389/fnbot.2021.660304.

[16] Shoulin Yin, Hang Li, Shahid Karim, and Yang Sun. ECID: Elliptic Curve Identity-based Blind Signature Scheme[J]. International Journal of Network Security, 23, No. 1, pp. 9-13, 2021.

[17] Lei W, Xiang Y, Wang Y, et al. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[18] Pitler E, Louis A, Nenkova A. Automatic sense prediction for implicit discourse relations in text[C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural

Language Processing of the AFNLP. Association for Computational Linguistics, 2009: 683-691.

[19] Lin Z, Kan M Y, Ng H T. Recognizing implicit discourse relations in the Penn Discourse Treebank C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association or Computational Linguistics, 2009: 343-351.

[20] Rutherford A, Xue N. Discovering implicit discourse , relations through brown cluster pair representation and coreferencepatterns[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014: 645-654.

[21] Braud C, Denis P. Comparing word representations for implicit discourse relation classification[C7//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 2201-2211.

[22] Ji Y, Eisenstein J. One vector is not enough: Entityaugmented distributional semantics for discourse relations[J]. arXiv:1411.6699, 2014.

[23] Zhang B, SuJ, Xiong D, et al. Shallow convolutional neural network for implicit discourse relation recognitionC7//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 2230-2235.

[24] Qin L, Zhang Z, Zhao H. A stacking gated neural architecture for implicit discourse relation classification [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2263-2270.

[25] Hong Y., Zhu S., Yan W., Yao J., Zhu Q., Zhou G. (2014) Expanding Native Training Data for Implicit Discourse Relation Classification. Social Media Processing. SMP 2014. Communications in Computer and Information Science, vol 489, pp. 67-75, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-45558-6_6

[26] Wu C, Chen Y, Huang Y. Bilingually-constrained synthetic data for implicit discourse relation recognition[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016: 2306-2312.

[27] Xu Y, Hong Y, Ruan H, et al. Using active learning to expand training data for implicit discourse relation recognition [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing,2018: 725-731.

[28] Ruan H, Hong Y, Sun Y, et al. Using WHY-typequestion-ans wer pairs to improve implicit causal relation recognition[C]//Proceedings of the International Conference on Asian Language Processing, 2019: 355-360.

[29] Devlin J, Chang M W, Lee K, et al. BERT: Pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[30] Wang M, Y Li, Y Zhang, et al. Spatio-temporal graph convolutional neural network for remaining useful life estimation of aircraft engines[J]. Aerospace Systems, 2020, 4(2).

[31] Baroud S, Chokri S, Belhaous S, et al. A Brief Review of Graph Convolutional Neural Network Based Learning for classifying remote sensing images[J]. Procedia Computer Science, 2021, 191(1):349-354.

[32] Subakan C, Ravanelli M, Cornell S, et al. Attention is All You Need in Speech Separation[J]. 2020. arXiv:2010.13154

[33] Shoulin Yin, Hang Li, Desheng Liu and Shahid Karim. Active Contour Modal Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation [J]. Multimedia Tools and Applications. Vol. 79, pp. 31049-31068, 2020.

[34] S. Yin and H. Li. Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582.

[35] Shahid Karim, Ye Zhang, Shoulin Yin*, Irfana Bibi. A Brief Review and Challenges of Object Detection in Optical Remote Sensing Imagery [J]. Multiagent and Grid Systems. 16(3), 227-243, 2020

[36] G. Xu, D. Sha, Y. Xu and X. Liao, "Dual-Transformer-Based DAB Converter With Wide ZVS Range for Wide Voltage Conversion Gain Application," in IEEE Transactions on Industrial Electronics, vol. 65, no. 4, pp. 3306-3316, April 2018, doi: 10.1109/TIE.2017.2756601.

[37] Yu M, Chen Z, Yao D, et al. Energy, exergy, economy analysis and multi-objective optimization of a novel cascade absorption heat transformer driven by low-level waste heat[J]. Energy Conversion and Management, 2020, 221:113162.

[38] Hendrycks D, Gimpel K. Gaussian error linear units (GeLUs)[J].arXiv:1606.08415, 2016.

[39] Yang Sun, Shoulin Yin, Hang Li, et al. GPOGC: Gaussian Pigeon-Oriented Graph Clustering Algorithm for Social Networks Cluster [J]. IEEE Access. Volume: 7, Page(s): 99254 - 99262, 03 July 2019.

[40] Shoulin Yin, Ye Zhang and Shahid Karim. Region search based on hybrid convolutional neural network in optical remote sensing images[J]. International Journal of Distributed Sensor Networks, Vol. 15, No. 5, 2019.

[41] Teng Lin, Hang Li and Shoulin Yin. Modified Pyramid Dual Tree Direction Filter-based Image De-noising via Curvature Scale and Non-local mean multi-Grade remnant multi-Grade Remnant Filter [J]. International Journal of

Communication Systems. v 31, n 16, November 10,pp. e.3486.1-e.3486.12, 2018.

[42] Shoulin Yin, Ye Zhang, Shahid Karim. Large Scale Remote Sensing Image Segmentation Based on Fuzzy Region Competition and Gaussian Mixture Model[J]. IEEE Access. volume 6, pp: 26069 - 26080, 2018.