TransS-Driven Joint Learning Architecture for Implicit Discourse Relation Recognition

Ruifang He[1,2,3], Jian Wang[1,2], Fengyu Guo[1,2]*, and Yugui Han[1,2]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China
[3]State Key Laboratory of Cognitive Intelligence, iFLYTEK, China
{rfhe,jian_wang,fengyuguo,yghan}@tju.edu.cn

Abstract

Implicit discourse relation recognition is a challenging task due to the lack of connectives as strong linguistic clues. Previous methods primarily encode two arguments separately or extract the specific interaction patterns for the task, which have not fully exploited the annotated relation signal. Therefore, we propose a novel TransS-driven joint learning architecture to address the issues. Specifically, based on the multi-level encoder, we 1) translate discourse relations in low-dimensional embedding space (called TransS), which could mine the latent geometric structure information of argument-relation instances; 2) further exploit the semantic features of arguments to assist discourse understanding; 3) jointly learn 1) and 2) to mutually reinforce each other to obtain the better argument representations, so as to improve the performance of the task. Extensive experimental results on the Penn Discourse TreeBank (PDTB) show that our model achieves competitive results against several state-of-the-art systems.

1 Introduction

Discourse relation describes how two adjacent text units (e.g., clauses, sentences, and larger sentence groups) are connected logically to one another. A discourse relation instance is usually defined as a connective taking two arguments (as *Arg*1 and *Arg*2, respectively). Implicit discourse relation recognition without explicit connectives (Pitler et al., 2009) is still a challenging problem of discourse analysis, which needs to infer the discourse relation from a specific context. It is beneficial to many downstream natural language processing (NLP) applications, such as machine translation (Meyer and Popescu-Belis, 2012) and text summarization (Gerani et al., 2014).

The existing neural network-based models have shown great success in recognizing implicit discourse relations. It mainly includes 1) Basic neural networks (Braud and Denis, 2015; Zhang et al., 2015; Liu et al., 2016) can learn the dense vector representations of discourse arguments, which can capture the semantic information to some extent. Further studies exploit different attention or memory mechanisms (Liu and Li, 2016; Zhang et al., 2016) to capture the critical information of argument pairs. 2) Complex neural models (Chen et al., 2016; Lei et al., 2017; Guo et al., 2018) utilize gated relevance networks or neural tensor networks to capture the deeper interactions between two discourse arguments. 3) Joint learning architectures (Qin et al., 2017; Bai and Zhao, 2018; Xu et al., 2019) exploit implicit connective cues, different granularity of text, or topic-level relevant information to improve the discourse relation prediction. However, these approaches still have the following drawbacks: 1) do not make full use of the annotated discourse relation signal to explore the argument-relation features; 2) neglect the extra information in the low-dimensional continuous embedding space, i.e., the direction or structure information of the vectors.

Notice that Translating Embeddings (TransE) is a method for the prediction of entities' missing relations in knowledge graphs. Bordes et al. (2013) model relations by interpreting them as translating operation not on the graph structure directly but in a learned low-dimensional embedding of the knowledge graph entities: if (h_e, l_e, t_e) holds, then the embedding of the tail entity t_e should be close to the embedding of the head entity h_e plus some vector that depends on the relation l_e. Similar to the entity relation extraction, our task aims to identify the semantic relations between two arguments (i.e., sentences).

Inspired by TransE, we design a new method

[*] Corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 139–148
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

(TransS), which translates discourse relations in sentence embedding spaces to mine the argument-relation features. Intuitively, these features reflect the latent geometric structure among the arguments and their discourse relation by performing the algebraic operation, and the argument-relation instances with the same discourse relation may have similar direction and position information in the embedding space. Therefore, we propose a novel TransS-driven joint learning neural network framework that leverages the latent geometric structure information of argument-relation instances, in addition to using the semantic features to improve the comprehension of discourse argument. Among them, we adopt a multi-level encoder to further enrich the argument representations, which could obtain the deeper semantics of discourse.

In summary, the main contributions of this paper are as follows:

- Propose a novel TransS-driven joint learning architecture, including the latent geometric structure information learning (GSL) and semantic feature learning (SFL);
- Design TransS approach to translate discourse relations in low-dimensional embedding space from the sentence-level perspective, which could induce the geometric structure of argument-relation instances to some extent;
- Employ the mutual reinforcing between the GSL and SFL to optimize the argument representations: 1) the GSL adopts its geometric structure clues to facilitate the SFL; 2) the SFL utilizes its semantic cues to improve the learning capability of GSL;
- The experimental results on the PDTB demonstrate the effectiveness of our model.

2 The Proposed Model

The implicit discourse relation recognition task is usually formalized as a classification problem. In this section, we give an overview of the TransS-driven joint learning framework, which consists of four parts: embedding layer, multi-level encoder, latent geometric structure learning, and semantic feature learning, as shown in Figure 1.

2.1 Embedding Layer

In order to model two discourse arguments with neural networks, we transform the one-hot representations of arguments and their discourse relation into the distributed representations. Formally, the embedding layer could be seen as a simple projection layer where the word embedding is achieved by lookup table operation according to the indexes. All words of two arguments *Arg*1, *Arg*2, and their relation will be mapped into low dimensional vector representations, which are taken as the input of our model.

2.2 Multi-level Encoder

To enrich the discourse argument representations, we exploit multi-level encoder shown in Figure 2 to learn the argument representations at the different levels. Particularly, the higher-level states of multi-level encoder could capture context-dependent aspects of words while the lower-level states could model aspects of syntax (Peters et al., 2018). The multi-level encoder is composed of stacked encoder layers.

2.2.1 Encoder Layer

Referring to the previous work, we implement the bidirectional LSTM (BiLSTM) neural network to model the argument sequences, which could preserve both the historical and future information in forward and reverse directions. Therefore, we can obtain two representations $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ at each time step t of the sequence. Then we concatenate them to get the intermediate state $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$.

Attention Controller. Due to the limitations of treating each word equally in the general representations, we use attention mechanism to point out the words particularly useful for our task. Let H be the matrix consisting of output vectors $[h_1, h_2, ..., h_n]$ of the last layer produced, where n is the length of the argument. The new representation $\tilde{h}$ of the argument is formed by a weighted sum of the output vectors:

$$M = tanh(H), \tag{1}$$

$$\alpha = softmax(w^T M), \tag{2}$$

$$\tilde{h} = H\alpha^T. \tag{3}$$

where $H \in \mathbb{R}^{n \times d}$, d is the dimension of word embedding, w is a parameter vector. Then we could obtain the argument representation with important information from Eq. (4) for the next step.

$$h^* = tanh(\tilde{h}) \tag{4}$$

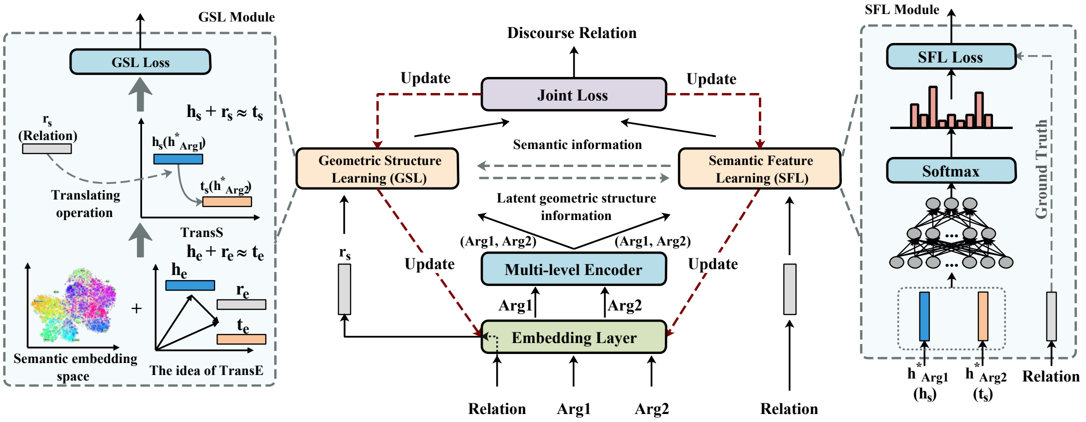


Figure 1: TransS-driven joint learning architecture of our proposed model.

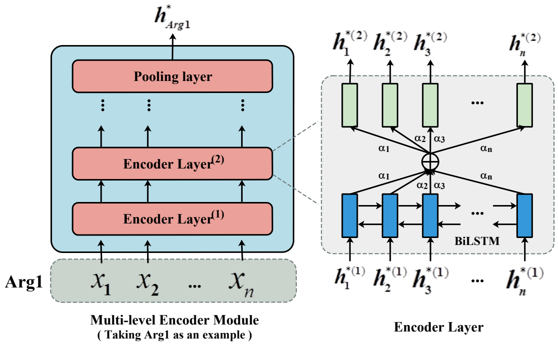

h^*_{Arg1}
Pooling layer
Encoder Layer$^{(2)}$
Encoder Layer$^{(1)}$
Arg1
x_1 x_2 ... x_n
Multi-level Encoder Module
(Taking Arg1 as an example)
$h_1^{*(2)}$ $h_2^{*(2)}$ $h_3^{*(2)}$ $h_n^{*(2)}$
α1 α2 α3 αn
BiLSTM
$h_1^{*(1)}$ $h_2^{*(1)}$ $h_3^{*(1)}$ ... $h_n^{*(1)}$
Encoder Layer

Figure 2: The illustration of multi-level encoder.

2.2.2 Pooling Layer

Finally, we can receive the overall argument representations by averaging pooling operation for the word embedding sequence, defined as:

$$h^*_{Arg} = \frac{1}{n}\sum_{i=1}^{n} h_i^{*(m)} \tag{5}$$

where h^*_{Arg} is the argument representation, $h_i^{*(m)}$ is the representation of the i-th word in the word embedding sequence of the m-th encoder layer, n is the number of words in an argument.

2.3 Latent Geometric Structure Learning

TransE, as a model for learning low-dimensional embeddings of entities, is to enforce the structure of embedding space in which different relations between entities of different types may be represented by translation (Bordes et al., 2013). Discourse relation recognition and entity relation extraction are similar to some extent. Intuitively, the argument-relation instances with the same discourse relation may also have similar direction and position information in embedding space. However, discourse argument embedding is a sentence-level representation, which is different from the reuse of entities in other sentences, and more diverse and complex than entity representation. Therefore, we design **TransS**, a method which models discourse relations by interpreting them as translations operating in the low-dimensional embedding space from the sentence perspective. Moreover, it could mine the latent geometric structure of argument-relation instances. Specifically, to define two arguments as head vector h_s and tail vector t_s respectively, their annotated relation signal as relation vector r_s, the latent geometric structure is reflected by $h_s + r_s \approx t_s$, their score function is defined as follows:

$$d_s(h_s, t_s) = ||h_s + r_s - t_s||_2^2. \tag{6}$$

where h_s, t_s denote the representations of $Arg1$ and $Arg2$ respectively; $r_s \in \mathbb{R}^d$ is the embedding of discourse relation and d is the dimension of word embedding.

GSL Loss. Under the framework of TransS, given a training set T of triplets (h_s, r_s, t_s) composed of two arguments $h_s, t_s \in V$ (the set of sentence vectors) and a relation $r_s \in R$ (the set of relation), our model would learn the embeddings of the words in arguments and the discourse relation. The GSL loss function is defined as:

$$\mathcal{L}_{GSL} = \sum_{(h_s,r_s,t_s)\in T} \sum_{(h'_s,r_s,t'_s)\in T'_{(h_s,r_s,t_s)}} [\gamma + d_s(h_s + r_s, t_s) - d_s(h'_s + r_s, t'_s)]_+ + \lambda_{GSL}\|\theta\|_2^2. \quad (7)$$

where $[\cdot]_+$ denotes the positive instances, $\gamma > 0$ is a margin hyper-parameter, and the set of negative triplets, constructed according to Eq.(8), in which the head or tail is replaced by a random argument vector (but not simultaneously). θ denotes the other parameters of the network. L_2 regularization is used to penalize the size of all parameters for preventing overfitting, weighted by λ_{GSL}.

$$\begin{aligned} T'_{(h_s,r_s,t_s)} = &\{(h'_s, r_s, t_s)|h'_s \in V\}\cup \\ &\{(h_s, r_s, t'_s)|t'_s \in V)\}. \end{aligned} \quad (8)$$

By optimizing the GSL loss, we could obtain the latent geometric structure information about argument-relation instances. Different from TransE, we could not directly utilize TransS to recognize discourse relations, for that each argument could not be reused in discourse. Therefore, we exploit TransS to mine the latent geometric structure information and further guide the semantic feature learning.

2.4 Semantic Feature Learning

The new argument representations (h^*_{Arg1}, h^*_{Arg2}) with latent geometric structure information learned by the GSL are as inputs of the semantic feature learning (SFL). The h^*_{Arg1}(i.e., h_s) and h^*_{Arg2}(i.e., t_s) are obtained from the multi-level encoder. We further stack a softmax layer upon the representations:

$$y = f(W_f \begin{bmatrix} h^*_{Arg1}, \\ h^*_{Arg2} \end{bmatrix} + b_f). \quad (9)$$

where f is the softmax function, $W_f \in \mathbb{R}^{C\times 2d}$, $b_f \in \mathbb{R}^C$ are the weights and bias term respectively, d denotes the dimension of word embedding and C denotes the number of relation classes.

SFL Loss. Under the framework of basic neural networks for our task, given training set T, two argument vectors h_s, t_s in the triplet (h_s, r_s, t_s) are concatenated to a new sentence vector during the training process, and then the generated vector is used for relation recognition. The SFL loss is a cross-entropy style shown as:

$$\mathcal{L}_{SFL} = -\sum_{j=1}^{C} y_j log(\hat{y_j}) \quad (10)$$

where y is the one-hot representation of the ground-truth relation; $\hat{y}$ is the predicted probabilities of relations; C is the number of relation class.

2.5 Joint Learning

After obtaining the new representations *Arg*1 as head vector h_s, *Arg*2 as tail vector t_s, and the relation vector r_s, our model is trained using joint learning mechanism. The goal of our model is to minimize the loss function (Eq.(11))

$$\mathcal{L} = \mathcal{L}_{GSL} + \lambda\mathcal{L}_{SFL}. \quad (11)$$

where, $\mathcal{L}_{GSL}$ and $\mathcal{L}_{SFL}$ are from Eq.(7) and (10), respectively; λ is the trade-off parameter controlling the balance between GSL and SFL.

Our model jointly learns the GSL and SFL to optimize the argument representations. On the one hand, the GSL maps the discourse relation between two arguments to the low-dimensional embedding space and obtains the vectors h_s, r_s, t_s with geometric structure information to constrain the SFL. On the other hand, the SFL alternately optimizes the discourse representations and provides the necessary semantic clues for geometric structure information mining. Generally, the GSL and SFL reinforce with each other, and finally get the better argument representations containing the semantics and the latent geometric structure information of argument-relation.

3 Experiments

3.1 Datasets

The PDTB 2.0, a large scale corpus annotated on 2,312 Wall Street Journal articles, is utilized for all experiments. It contains three hierarchies: Level-1 Class, Level-2 Type, and Level-3 Subtype. We focus on the first level, which contains four classes: Comparison (Comp.), Contingency (Cont.), Expansion (Exp.), and Temporal (Temp.). As (Rutherford and Xue, 2014), we use Sections 2-21 as the training set, Section 22 as the development set, Section 23 as the test set.

Relation	Train	Dev	Test
Comp.	1945	196	152
Cont.	3242	284	272
Exp.	6794	646	546
Temp.	709	61	79
Total	12690	1187	1049

Table 1: The statistical distribution of PDTB.

3.2 Experimental Settings

All the arguments are padded at the same length of 100. Word embedding is randomly initialized by uniformly distributed samples [-0.1, 0.1] with 300-dimension. The learning rate is set to 0.001, the batch size is 128, and the number of iteration is 100. For the GSL, the margin of loss is set to 0.5, the trade-off parameter λ in Eq.(11) is set to 1.0, and we use L_2 distance as dissimilarity; For the SFL, the sizes of the input and the hidden layer of the BiLSTMs are both 300; we choose three encoder layers, and set the dimension of pre-trained embeddings from ELMo (Peters et al., 2018) to 300.

3.3 The Comparison Models

3.3.1 The State-of-the-art Systems

To validate the effectiveness of our model, we select some state-of-the-art systems from the following three aspects to compare with our model:

• **Discourse Argument Representation**

1) **Ji2015**: Ji and Eisenstein (2015) computed distributed representations for each discourse argument by composition up the syntactic parse tree.

2) **Zhang2015**: Zhang et al. (2015) proposed pure neural networks with three different pooling operations to learn shallow representations in tasks.

3) **Liu2016a**: Liu and Li (2016) combined attention mechanism and external memory to focus on specific words that helps determine discourse relations.

4) **Lan2017**: Lan et al. (2017) designed an attention-based neural network for learning discourse argument representations and a multi-task framework for learning knowledge from annotated and unannotated corpora.

• **Complex Neural Models**

5) **Chen2016**: Chen et al. (2016) adopted a gated relevance network to capture interaction information between two arguments to enhance relation recognition.

6) **Qin2016**: Qin et al. (2016a) adopted context-aware character-enhanced embeddings to address implicit discourse relation recognition task.

7) **Lei2017**: Lei et al. (2017) devised the Simple Word Interaction Model (SWIM) to learn the interactions between word pairs.

8) **Dai2018**: Dai and Huang (2018) modeled interdependencies between discourse units as well as discourse relation continuity and patterns, and predict a sequence of discourse relations in a paragraph.

• **Joint Learning**

9) **Liu2016b**: Liu et al. (2016) designed related discourse classification tasks specific to a corpus, and proposed a novel Convolutional Neural Network embedded multi-task learning system to synthesize these tasks by learning both unique and shared representations for each task.

10) **Bai2018**: Bai and Zhao (2018) employed different grained text representations, including character, subword, word, sentence, and sentence pair levels, and transfered the knowledge from the implicit connectives to support discourse relation prediction.

3.3.2 The Ablation Methods

In order to validate the effectiveness of each component of our model, we present the following ablation methods:

- **Baseline (Including SFL)** We use three encoder layers to encode the argument pairs separately, then concatenate them together, and feed them to the SFL module for relation recognition.

- **+GSL** We encode two arguments based on the Baseline, and then feed them into GSL and SFL modules, respectively. Finally, we use the two modules to help recognize the discourse relation.

- **+ELMo** We utilize the Baseline to receive the argument representations, and then we use the pre-trained ELMo vector to enhance the argument representations. Finally, we feed them to the SFL module for relation recognition.

- **+GSL & ELMo (Ours)** We feed the two argument representations, encoded by the Baseline and enhanced by the pre-trained ELMo vector, into GSL and SFL modules, respectively. And then, we utilize the integrated representation to recognize the discourse relation.

3.4 Results and Discussion

Consistent with previous studies, we choose F_1 score and accuracy as evaluation metrics. For binary classification, the result is computed by F_1 score, and for 4-way classification, the result is computed by macro average F_1 score.

Model	Comp.	Cont.	Exp.	Temp.	4-way	Acc.
Ji2015	35.93	52.78	-	27.63	-	-
Zhang2015	33.22	52.04	69.59	30.54	-	-
Liu2016a	32.13	46.09	69.88	31.82	44.98	57.27
Lan2017	40.73	**58.96**	72.47	38.50	47.80	57.39
Chen2016	40.17	54.76	-	31.32	-	-
Qin2016	38.67	54.91	**80.66**	32.76	-	-
Lei2017	40.47	55.36	69.50	35.34	46.46	-
Dai2018	37.72	49.39	67.45	**40.70**	48.82	59.75
Liu2016b	39.86	54.48	70.43	38.84	46.29	57.57
Bai2018	47.85	54.47	70.60	36.87	51.06	-
Ours	**47.98**	55.62	69.37	38.94	**51.24**	**59.94**

Table 2: F_1 score (%) and Accuracy(Acc., %) of different comparison models on binary and 4-way classification.

Model	Comp.	Cont.	Exp.	Temp.	4-way	Acc.
Baseline	32.32	49.53	65.91	34.86	46.46	54.02
+ GSL	44.88	53.17	67.91	37.38	48.91	57.65
+ ELMo	46.85	54.57	68.44	38.71	50.07	58.89
+ GSL & ELMo (Ours)	47.98	55.62	69.37	38.94	51.24	59.94

Table 3: F_1 score (%) and Accuracy(Acc., %) of ablation models on binary and 4-way classification.

3.4.1 Comparison with the state-of-the-art Systems

Table 2 shows the results of the compared state-of-the-art systems on binary and 4-way classification. We could make the following observations:

- Overall, i) our model achieves state-of-the-art performance, i.e., the F_1 score and accuracy are 51.24% and 59.94% on the 4-way classification, respectively; ii) the results of binary classification are keeping a similar tendency with the 4-way classification. In particular, our model gains the best F_1 score on Comparison relation. The main reasons may be that the instances with different discourse relations have different directions and position (geometric structure) features in the low-dimensional continuous embedding space, and the Comparison instances have more obvious indicative structure features.

- Comparing our model with Chen2016 and Lei2017, the F_1 scores of our model are higher than those of the latter two. It proves that our model is better than the two methods only considering the content interactions, since we jointly leverage the geometric structure information and the semantic information of the argument-relation instances to obtain deeper interactions.

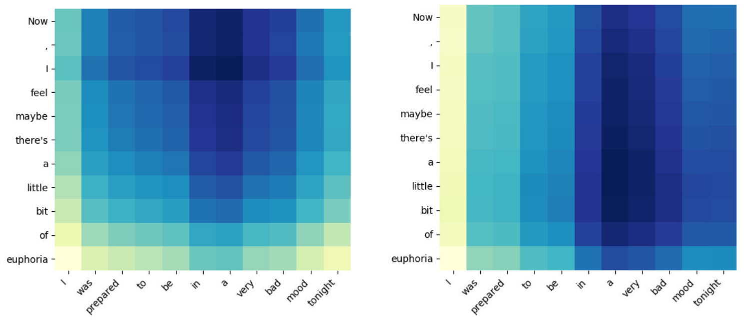
- In the comparison models, Bai2018 with joint learning framework achieves the best performance, which illustrates that jointly utilizing the discourse relation and the implicit connectives are helpful to the task. Moreover, the performance of our model is better than that of Bai2018. It not only indicates that the effectiveness of joint learning, but also proves considering the geometric structure is beneficial to our task.

3.4.2 Ablation Models

For the ablation models, we can make the observations from Table 3:

Overall:1) Our model gains state-of-the-art performance than that of the other ablation models. This demonstrates that the geometric structure information could enrich the argument representation and promote implicit discourse relation recognition. 2) All models have a higher F_1 values on the Expansion relation than those of the other relations. The unbalanced data may cause that.

GSL: The F_1 score of our model using the GSL module is 48.91%, higher than the performance of Baseline. In addition, compared with ELMo,



(a) without geometric structure features. (b) with geometric structure features.

Figure 3: Visualization of the interaction information of argument representation.

although the performance of GSL does not exceed ELMo's, GSL obtain comparable results. This manifests that the two modules (GSL and SFL) could reinforce with each other, which utilizes the geometric structure information by the algebraic operation. Moreover, we exploit the geometric structure clues to augment the semantic understanding of discourse from a new aspect, which is different from the ELMo only focusing on the semantic information of the text itself.

ELMo: The third row of Table 3 is the result of our model, which only uses the pre-trained ELMo vector to enhance argument representations. The F_1 score and accuracy are 50.07% and 58.89%, respectively, which achieve 3.61% and 4.87% improvements than those of the Baseline. It verifies that ELMo, as pre-trained contextualized word embeddings, could contain more contextual information.

GSL & ELMo: Compared with ELMo, GSL & ELMo gains better performance, which demonstrates that inducing spatial geometry structure information based on argument enhancement could understand the semantics of discourse better.

3.4.3 Impact of TransS

To illustrate the effectiveness of the latent geometric structure information of argument-relation instances gotten by TransS, we visualize the heat maps of the interaction information of argument representations shown in Figure3. Every word comes with various background colors. The darker patches denote the correlations of word pairs are higher. The example of Comparison relation is listed below:

Arg1: I was prepared to be in a very bad mood tonight.

Arg2: Now, I feel maybe there's a little bit of euphoria.

From the semantics of perspective, this example could be identified as Comparison or Temporal relation. Since argument pairs may have distinct distinguishing features in geometric space, we could consider the geometric structure of argument pairs to help identify the discourse relation. We can obtain the following observations:

- Seen from Figure3(a), without introducing geometric structure information, the model has a high correlation around the word "Now" which might indicate the Temporal relation directly. This demonstrates that only considering the semantic information of arguments may suffer from issues such as polysemy, ambiguity, as well as fuzziness.

- Figure3(b) shows the result of the interaction information of argument representations, which introduces the GSL. From the results, we can see that the model has a high correlation around the word "little" and "very" with the comparative information. The possible reason is that our model utilizing GSL shifts the higher attention from the word "Now" with Temporal information to the word pairs (little, very), (euphoria, bad) and (euphoria, mood) with Comparison relation. Our model with GSL introduces the geometric structure information and jointly utilizes these features and semantic information to help identify the discourse relation.

3.4.4 Impact of Encoder Layer Number

In order to illustrate the impact of the encoder layer number, we select different sizes of encoder layer

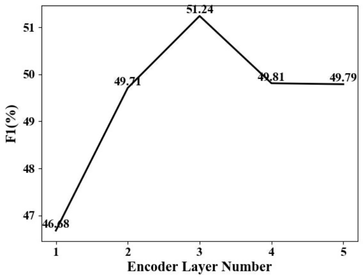


Figure 4: The effect of encoder layers' number.

as comparison experiments on the 4-way classification. Figure 4 shows that the F_1 scores are increasing until three encoder layers. And when the size of the encoder layer is four or five, the performance of our model is decreasing obviously.

With the increasing of the number of encoder layers, the model could capture the richer semantic information. However, the results imply that with the more encoder layers considered, the model could incur the over-fitting problem due to adding more parameters. Therefore, we adopt three encoder layers to encode the arguments as our Baseline in section 3.3.

4 Related Work

Neural network-based models have shown great effectiveness in implicit discourse relation recognition. We give the analysis of mainly relevant work:

4.1 Discourse Argument Representation

Proper argument representation is a core factor of our task. Most previous researches encode arguments as dense and continuous representation based on various neural networks, from basic neural networks (such as CNN, RNN) to complex neural networks (Zhang et al., 2015; Qin et al., 2016b; Rutherford et al., 2016). Some studies adopt different attention or memory mechanisms to catch the emphasis on discourse arguments (Mnih et al., 2014; Liu and Li, 2016; Zhang et al., 2016). Li et al. (2016) exploit the hierarchical attention to capture the focus of different granularities. Zhang et al. (2016) build upon a semantic memory to store knowledge in the distributed fashion for the task. However, these models have only considered the two arguments independently without the interaction information.

4.2 Argument Pair Interactions

Further studies tend to discover more semantic interactions between two arguments by complex neural networks (Qin et al., 2016c; Cai and Zhao, 2017; Lan et al., 2017; Guo et al., 2018). Chen et al. (2016) develop a novel gated relevance network to capture semantic interactions between arguments. Lei et al. (2017) conduct word pair interaction score to capture both linear and quadratic relation for argument representation. However, these methods utilize the pre-trained embeddings for mining the interaction features and ignore the geometric structure information entailed in discourse arguments and their relation.

4.3 Joint Learning Perspective

Recently, some researches adopt joint learning framework to capture more discourse clues for the task. Bai and Zhao (2018) jointly predict connectives and relations, assuming the shared parameters of the deep learning models. Xu et al. (2019) propose a topic tensor network (TTN) to model the sentence-level interactions and topic-level relevance among arguments for this task. However, few studies model discourse relations by translating them in the low-dimensional embedding space as we do in this work.

TransE effectively maps the relation to the embedding space of entities by performing the algebraic operation. Bordes et al. (2013) model entity relations by interpreting them as translating operation in the low-dimensional embedding of the entities. Inspired by TransE, we design a TransS method to mine the latent geometric structure information, which could enhance the argument representations for promoting discourse relation recognition. To our knowledge, this is the first attempt to mine the latent geometric structure of argument-relation. Meanwhile, the embeddings of argument and relation by TransS could be used to the other high-level NLP tasks.

5 Conclusion

In this paper, we propose a novel TransS-driven joint learning neural network framework by optimizing the discourse argument representations to improve implicit discourse relation recognition. We interpret the discourse relations as translation in low-dimensional embedding space, which reflects the geometric structure of argument-relation, and also can obtain the richer argument representations based on the multi-level encoder. Different from the conventional approaches only considering the semantic features, we jointly leverage the latent geometric structure information and the semantic features to optimize the argument representations, which could improve the semantic understanding of discourse. Experimental results on the PDTB show the effectiveness of our model.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Our work is supported by the National Key R&D Program of China (2019YFC1521200), the National Natural Science Foundation of China (61976154,U1736103), the Tianjin Natural Science Foundation (18JCY-BJC15500), and the Foundation of State Key Laboratory of Cognitive Intelligence, iFLYTEK (CIOS-20190001).

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th COLING*, pages 571–583.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th NIPS*, pages 2787–2795.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 EMNLP*, pages 2201–2211.

Deng Cai and Hai Zhao. 2017. Pair-aware neural sentence modeling for implicit discourse relation classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 458–466. Springer.

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 ACL*, 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 NAACL*, pages 141–151.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 EMNLP*, pages 1602–1613.

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 COLING*, pages 547–558.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributional 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 EMNLP*, pages 1299–1308.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *Proceedings of the 26th IJCAI*, pages 4026–4032.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 EMNLP*, pages 362–371.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 EMNLP*, 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 30th AAAI*, pages 2750–2756.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the 13th EACL*, pages 129–138.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Proceedings of the 27th NIPS*, pages 2204–2212.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 NAACL*, pages 2227–2237.

Emily Pitler, Annie Louis, and Ani and Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the AFNLP*, pages 683–691.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016a. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In *Proceedings of the 26th COLING*, pages 1914–1924.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016b. Shallow discourse parsing using convolutional neural network. In *CoNLL Shared Task*, pages 70–77.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016c. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 EMNLP*, 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 ACL*, pages 1006–1017.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th EACL*, pages 645–654.

Attapol T Rutherford, Vera Demberg, and Nianwen Xue. 2016. Neural network models for implicit discourse relation classification in english and chinese without surface features. *arXiv preprint arXiv:1606.01990*.

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 ACL*, pages 608–618.

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 EMNLP*, pages 2230–2235.

Biao Zhang, Deyi Xiong, and Jinsong Su. 2016. Neural discourse relation recognition with semantic memory. *arXiv preprint arXiv:1603.03873*.