Topic Tensor Network for Implicit Discourse Relation Recognition in Chinese

Sheng Xu, Peifeng Li, Fang Kong, Qiaoming Zhu and **Guodong Zhou**
Institute of Artificial Intelligence, School of Computer Science
and Technology, Soochow University, China
sxu@stu.suda.edu.cn;{pfli,kongfang,qmzhu,gdzhou}@suda.edu.cn

Abstract

In the literature, most of the previous studies on English implicit discourse relation recognition only use sentence-level representations, which cannot provide enough semantic information in Chinese due to its unique paratactic characteristics. In this paper, we propose a topic tensor network to recognize Chinese implicit discourse relations with both sentence-level and topic-level representations. In particular, besides encoding arguments (discourse units) using a gated convolutional network to obtain sentence-level representations, we train a simplified topic model to infer the latent topic-level representations. Moreover, we feed the two pairs of representations to two factored tensor networks, respectively, to capture both the sentence-level interactions and topic-level relevance using multi-slice tensors. Experimentation on CDTB, a Chinese discourse corpus, shows that our proposed model significantly outperforms several state-of-the-art baselines in both micro and macro F1-scores.

1 Introduction

As a critical component of discourse parsing, discourse relation recognition focuses on determining how two adjacent discourse units (e.g., clauses, sentences, and sentence groups), called arguments, semantically connect to one another. Obviously, identifying discourse relations can help many downstream NLP applications, such as automatic summarization, information extraction and question answering.

In principle, the discourse connectives between two arguments are important for recognizing the relationship between them. For explicit discourse relation recognition where the discourse connectives explicitly exist in the text, a simple frequency-based mapping table can achieve high performance due to the critical role of a connective in determining the discourse relations (Xue et al., 2016). For implicit discourse relation recognition, it is much more challenging due to missing an exact connective and it normally depends on the understanding of the whole text (Pitler et al., 2009).

This paper focuses on recognizing implicit discourse relations in Chinese. In contrast to English, which is a hypotactic language (formal cohesion), Chinese is a paratactic language (semantic cohesion) that tends to pro-drop clause connectives. Our statistics indicate that the implicit relations in the Chinese CDTB corpus account for 75.2%, while the proportion in the English PDTB corpus declines to only 40%. Hence, recognizing implicit discourse relations in Chinese becomes more crucial than in English.

In the literature, most of previous studies focused on English, with only a few on Chinese. Compared with traditional feature-based methods (Pitler et al., 2009; Lin et al., 2009; Wang et al., 2017; Kong and Zhou, 2017) that directly rely on feature engineering, recent neural network models (Liu et al., 2017; Qin et al., 2017; Guo et al., 2018; Bai and Zhao, 2018) can capture deeper semantic cues and learn better representations (Zhang et al., 2015). In particular, most neural network-based methods encode arguments using variants of Bi-LSTM or CNN (Qin et al., 2016; Guo et al., 2018) and propose various models (e.g., the gated relevance network, the encoder-decoder model, and interactive attention) to measure the semantic relevance (Chen et al., 2016; Cianflone and Kosseim, 2018; Guo et al., 2018)

Due to the large differences between the hypotactic English language and the paratactic Chinese language, English-based models, which rely heavily on sentence-level representations, may not function well on Chinese. Due to its paratactic nature, Chinese is flooded with a broad range of flexible sentence structures and semantic cohesion, such as ellipses, references, substitutions, and con-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 608–618
Florence, Italy, July 28 - August 2, 2019.

junctions. Therefore, Chinese discourse parsing relies heavily on the deep semantics of arguments, especially topic continuity (Lei et al., 2018). In many cases, considering only the sentence-level representation is not enough for Chinese implicit discourse relation recognition, and we need various semantic clues beyond the sentence-level, e.g., at the topic level. Take the following two arguments as examples:

[一九九一年至一九九五年，中国的对外开放以高速向前推进 (*From 1991 to 1995, China's opening was moving forward at a high speed*)]$_{Arg1}$ [国民经济更加广泛地参与国际分工与国际交换，中外经济技术合作与交流已渗入到中国经济生活的各个领域 (*the national economy is more widely involved in the international division of labor and international exchange, and the economic and technological cooperation and exchanges between China and foreign countries had penetrated into various fields of China's economic life*)]$_{Arg2}$

Although there is an *Elaboration* relation between the above two arguments, it is difficult to obtain sufficient information for identifying this potential association by directly matching the words in $Arg1$ (e.g., "speed" and "moving") and those in $Arg2$ (e.g., "economic" and "exchanges"). To identify their *Elaboration* relation, the most crucial clue may be the fact that they belong to the same topic, i.e., China's opening is an international economic event. Therefore, it is critical for implicit discourse relation recognition to capture such topic information as an important clue.

In this paper, we propose a Topic Tensor Network (TTN) to recognize implicit discourse relations in Chinese using both sentence-level and topic-level representations. First, we introduce a GCN-based (Gated Convolutional Network) encoder to learn the sentence-level representations. Then, we train a Simplified Topic Model (STM) to infer the latent topic-level representations to provide additional semantic clues. Finally, we feed the two pairs of representations to two Factored Tensor Networks (FTNs) to model both the sentence-level interactions and topic-level relevance using multi-slice tensors. We summarize the contributions of our work as follows:

- Compared with previous works that were focused on sentence-level representations, we incorporate additional topic-level representations to capture the deep semantic interactions among arguments.
- We introduce the simplified topic model STM to infer the latent topic-level representations and employ such topic-level relevance to recognize Chinese implicit discourse relations.
- We propose the factored tensor network FTN to model the complex semantic interactions, and it has the advantage of significantly reducing the complexity of the original model (Guo et al., 2018).

2 Related Work

Most previous studies evaluated their models on PDTB (Prasad et al., 2008) and RST-DT (Carlson et al., 2003), which are two English discourse corpora that were available up to now. PDTB is the largest English discourse corpus with 2312 annotated documents from Wall Street Journal using the PTB-style predicate-argument structure. RST-DT is another popular English discourse corpus, which annotates 385 documents from Wall Street Journal using the RST tree scheme.

Basically, previous studies can be categorized into traditional models that focus on linguistically informed features (Pitler et al., 2009; Lin et al., 2009; Feng and Hirst, 2014; Wang et al., 2017), and neural network methods (Liu and Li, 2016; Chen et al., 2016; Guo et al., 2018; Bai and Zhao, 2018). Especially, Zhou et al., (2010) attempted to predict implicit connectives. Qin et al. (2017), Shi et al. (2017) and Xu et al. (2018) attempted to leverage explicit examples for data augmentation. Other studies resorted to unlabeled data to perform multi-task or unsupervised learning (Liu et al., 2016; Lan et al., 2017).

Since discourse relation recognition is essentially a classification problem, what those neural network methods need to consider is how to model the arguments and how to incorporate their semantic interactions. From this regard, most of them focused on improving representations or incorporating the complex interactions. Bai and Zhao (2018) proposed a deep enhanced representation to represent arguments at the character, subword, word, and sentence levels. Chen et al. (2016) introduced a gated relevance network to model both the linear and nonlinear correlations between two arguments. Guo et al. (2018) used a neural tensor network to capture the interactive features with

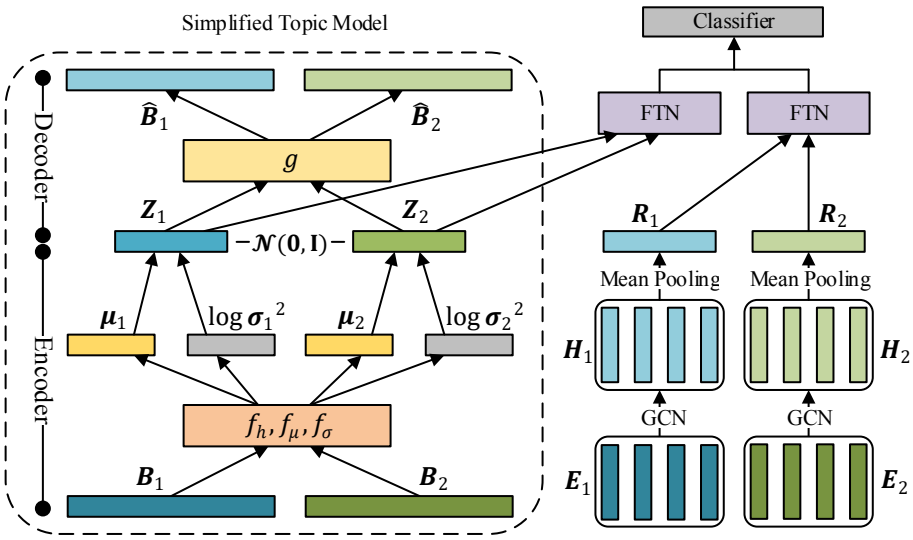

Figure 1: The overall framework of our Topic Tensor Network.

a multi-slice tensor. Among others, Qin et al. (2017) applied an adversarial method to transfer the discriminability of connectives to implicit features through competition, while Xu et al. (2018) expanded the training set by cooperating active learning with explicit-to-implicit relation transformation.

In comparison, previous studies on Chinese implicit discourse relation recognition were mainly carried out on CDTB (Li et al., 2014) and CDTB-ZX (Zhou and Xue, 2015). CDTB includes 500 newswire documents annotated with a connective-driven dependency tree scheme, while CDTB-ZX only contains 164 documents from Xinhua Newswire annotated with PDTB-style discourse relations.

Basically, most of the previous studies followed the English studies. Kong and Zhou (2017) constructed an end-to-end Chinese discourse parser, which used contextual features, lexical features and dependency tree features to recognize discourse relations with a maximum entropy classifier. Rönnqvist et al. (2017) proposed a Bi-LSTM model with attention mechanism to link two arguments by inserting special labels. Liu et al. (2017) provided a memory augmented attention model that used memory slots to store the interactions between two input arguments.

3 Topic Tensor Network for Implicit Discourse Relation Recognition

In this section, we describe our topic tensor network TTN with the overall architecture as shown in Figure 1. TTN has four major components: (1) a simplified topic model (STM) to infer the latent topic distributions of arguments as topic-level representations; (2) a GCN-based encoder to generate sentence-level representations; (3) two factored tensor networks (FTNs) to jointly model the sentence-level interactions and the topic-level relevance; and (4) an MLP classifier, which produces the final discourse relation labels.

In particular, the GCN-based encoder extracts hierarchical features from the long text of arguments by stacking multiple gated convolution layers, and fully represents the sentence-level semantic information. STM provides additional topic information for the MLP classifier to recognize discourse relations at a higher level. On this basis, the two pairs of representations are fed into two FTNs, respectively, which use multi-slice tensors to jointly model the sentence-level interactions and the topic-level relevance. Compared with the neural tensor network used in Guo et al. (2018), our FTN greatly reduces the computational complexity due to the tensor factorization. Hence, we can set more tensor slices to capture more complex interaction features.

Formally, the word sequence $\boldsymbol{E}_k = \{\boldsymbol{w}^1, \boldsymbol{w}^2, ..., \boldsymbol{w}^L\}$ and the BoW (Bag-of-Words) representation $\boldsymbol{B}_k \in \mathbb{R}^V$ of arguments are the input of our model, where L is the sequence length and V is the vocabulary size. Each word $\boldsymbol{w}^i$ in an argument is represented as the combination of its word embedding $\boldsymbol{e}^i$ and POS (Part-Of-Speech) embedding $\boldsymbol{p}^i$. The two word

sequences $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ of the two arguments are fed into the GCN-based encoder to obtain the sentence-level representations, and the BoW representations $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are sent to STM to infer the latent topic-level representations. On this basis, two FTNs are applied to capture the interactive features between two arguments based on the above representations. Finally, the MLP classifier concatenates all of the features produced by FTNs to predict the discourse relation label y.

3.1 Simplified Topic Model on Topic-level Representation

Similar to the LDA-style topic models, we believe that there is an association between the word distribution $\boldsymbol{B}_k$ of an argument and its topic distribution $\boldsymbol{Z}_k$. For each $\boldsymbol{B}_k$, we can infer a latent topic distribution $\boldsymbol{Z}_k \in \mathbb{R}^K$ through our topic model, where K denotes the number of topics. Inspired by the Neural Topic Model (NTM) (Zeng et al., 2018; Miao et al., 2016), we propose a simplified topic model STM based on the Variational AutoEncoder (VAE) (Kingma and Welling, 2013). Unlike NTM, our model does not attempt to reconstruct the document during the decoding phase, and it only restores the word distributions. Although STM cannot learn the semantic word embeddings, it significantly reduces the training parameters to perform unsupervised training on the discourse corpus with a small sample size.

Similar to NTM, we can interpret our STM as a VAE: a neural network encoder $p(\boldsymbol{Z}|\boldsymbol{B})$ first compresses the BoW representation $\boldsymbol{B}_k$ into a continuous hidden vector $\boldsymbol{Z}_k$, and then an MLP decoder $g(\boldsymbol{Z})$ restores $\boldsymbol{Z}_k$ to $\boldsymbol{B}_k$. Since STM is an unsupervised model, we can only use the existing BoW representation $\boldsymbol{B}_k$ to learn the latent topic distribution $\boldsymbol{Z}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. The inference network $p(\boldsymbol{Z}|\boldsymbol{B})$ is defined as follows:

$$\boldsymbol{\mu} = f_\mu(f_h(\boldsymbol{B})) \tag{1}$$

$$\log \boldsymbol{\sigma}^2 = f_\sigma(f_h(\boldsymbol{B})) \tag{2}$$

where $f_h(\cdot)$ is a single layer neural network with ReLU as the activation function, and $f_\mu(\cdot), f_\sigma(\cdot)$ are simple linear transformations. For the BoW representation $\boldsymbol{B}_k$ of the argument, the inference network generates its own parameters $\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2$ that parameterize the normal distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$, and we can further sample the latent topic distribution $\boldsymbol{Z}_k$ corresponding to the argument. To reduce the variance in the stochastic estimation, we follow (Rezende et al., 2014) to sample $\boldsymbol{Z}$ by the reparametric method and sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as follows:

$$\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \cdot \boldsymbol{\sigma} \tag{3}$$

We hope that our STM can reconstruct the original input $\boldsymbol{B}$ as much as possible using the topic distribution $\boldsymbol{Z}$ while adding Gaussian noise to the result generated by the encoder to increase the robustness of the decoder. Therefore, the loss function of STM is defined as follows:

$$\mathcal{L}_{STM} = \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{B})}[-\log q(\boldsymbol{B}|\boldsymbol{Z})] + KL(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{B})) \tag{4}$$

where $q(\boldsymbol{Z})$ is a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. It is worth mentioning that reducing the reconstruction loss can make the decoder have the generative ability. We calculate the reconstruction loss by calculating the binary cross entropy between the BoW representation $\boldsymbol{B}_k$ and $\widehat{\boldsymbol{B}}_k$ reconstructed by the decoder. Since decreasing the KL (Kullback-Leibler) divergence makes all $p(\boldsymbol{Z}|\boldsymbol{B})$ approximate the standard normal distribution, the noise can be prevented from being zero with the result as follows.

$$KL(q(\boldsymbol{Z})\|p(\boldsymbol{Z}|\boldsymbol{B})) = \frac{1}{2}(-\log \boldsymbol{\sigma}^2 + \boldsymbol{\mu}^2 + \boldsymbol{\sigma}^2 - 1) \tag{5}$$

Given the BoW representation $\boldsymbol{B}_k$, our STM can infer its latent topic distribution $\boldsymbol{Z}_k$ to provide topic-level representations.

3.2 GCN-based Encoder on Sentence-level Representation

Most previous studies used Bi-LSTM or 1D CNN to encode input sequences. However, CNN lacks visibility when capturing global information due to its limited view of the convolution kernel, while Bi-LSTM training is time-consuming due to its cyclic structure, especially for long texts, such as arguments. To address the above issues, Dauphin et al. (2017) proposed a Gated Convolutional Network (GCN) to extract hierarchical features from long texts by stacking multiple gated convolutional layers and mitigate the vanishing gradient problem by using gate units. In this paper, we choose GCN as our text encoder.

National Institute of Child Health and Human Development (2000) found that when readers repeatedly read text in detail with specific learning aims, they could improve not only their reading fluency, but also their comprehension of the text. Following He et al. (2016), we introduce the residual into GCN by adding the input of each layer

to its output so that the original input information can be passed to the back layers. Specifically, for the input sequence with N words $\boldsymbol{E} \in \mathbb{R}^{N \times D}$, where D is the sum of the size of the word embedding and POS embedding, each gated convolutional layer h_l is computed as follows:

$$h_l(\boldsymbol{X}) = (\boldsymbol{X} \cdot \boldsymbol{W} + \boldsymbol{b}) \otimes \sigma(\boldsymbol{X} \cdot \boldsymbol{V} + \boldsymbol{c}) + \boldsymbol{X} \quad (6)$$

where $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ is the input of layer h_l (either the input sequence $\boldsymbol{E}$ or the outputs of previous layers), $\boldsymbol{W} \in \mathbb{R}^{C \times D \times D}, \boldsymbol{b} \in \mathbb{R}^{D}, \boldsymbol{V} \in \mathbb{R}^{C \times D \times D}, \boldsymbol{c} \in \mathbb{R}^{D}$ are model parameters, and C is the size of the convolution kernel. $\sigma(\cdot)$ is the sigmoid function and $\otimes$ is the element-wise product between matrices. After stacking L layers on top of the input, we can obtain the semantic representation sequence of the argument $\boldsymbol{H} = h_L \circ ... \circ h_1(E) \in \mathbb{R}^{N \times D}$. Finally, the Mean Pooling operation is performed to obtain the respective argument representations on the sequences $\boldsymbol{H}_1 = \{h^1_{L1}, ..., h^N_{L1}\}$ and $\boldsymbol{H}_2 = \{h^1_{L2}, ..., h^N_{L2}\}$ corresponding to the two arguments:

$$\boldsymbol{R}_1 = \frac{1}{N}\sum_{i=1}^{N} h^i_{L1}, \quad \boldsymbol{R}_2 = \frac{1}{N}\sum_{i=1}^{N} h^i_{L2} \quad (7)$$

As a result, in the GCN-based encoder, we stack multiple gated convolution layers with the residual structure to learn the sentence-level representations, which can take advantage of the parallel computing of convolutional networks, and also control the flow of information through the gate units similar to LSTM.

3.3 Factored Tensor Network on Joint Representations

Traditional methods for modeling the semantic relevance between two arguments capture the linear and nonlinear interactions using various text matching models, such as Bilinear model (Jenatton et al., 2012) and Single Layer Network (Collobert and Weston, 2008). Based on these methods, Socher et al. (2013) proposed a Neural Tensor Network (NTN) to combine the advantages of these two models and showed the ability of the tensor to model complex informative interactions in knowledge graphs.

Following Guo et al. (2018), we use two NTNs to capture the interactive features between the semantic representations $\boldsymbol{R}_1, \boldsymbol{R}_2$, and between the topic distributions $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ as follows:

$$T(\boldsymbol{x}, \boldsymbol{y}) = f_n\left(\boldsymbol{x}^\top \boldsymbol{M}^{[1:m]} \boldsymbol{y} + \boldsymbol{U} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} + \boldsymbol{s}\right) \quad (8)$$

where $f_n(\cdot)$ is a standard nonlinear function, $\boldsymbol{M} \in \mathbb{R}^{d \times d \times m}$ is a 3rd-order transformation tensor, $\boldsymbol{U} \in \mathbb{R}^{m \times 2d}$ and $\boldsymbol{s} \in \mathbb{R}^{m}$ are parameters. The tensor product $\boldsymbol{x}^\top \boldsymbol{M}^{[1:m]} \boldsymbol{y}$ results in a vector $\boldsymbol{c} \in \mathbb{R}^{m}$, where each entry is computed by slice i of the tensor $\boldsymbol{M}$ as $\boldsymbol{c}_i = \boldsymbol{x}^\top \boldsymbol{M}^{[i]} \boldsymbol{y}$, and it is equivalent to including m Bilinear models that simultaneously capture multiple linear interactions between vectors. However, it increases the parameters and the computational complexity of the model; therefore, we adopt tensor factorization (Pei et al., 2014), which uses two low rank matrices to approximate each tensor slice $\boldsymbol{M}^{[i]}$, as follows:

$$\boldsymbol{M}^{[i]} \Rightarrow \boldsymbol{J}^{[i]} \boldsymbol{K}^{[i]} \quad (9)$$

where $\boldsymbol{J}^{[i]} \in \mathbb{R}^{d \times r}$, $\boldsymbol{K}^{[i]} \in \mathbb{R}^{r \times d}$ and $r \ll d$.

We named our model FTN (Factored Tensor Network). Compared with the original NTN (Guo et al., 2018), our FTN greatly reduces the number of parameters. Hence, it can set more tensor slices and make the training process easier. In particular, for semantic representations $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathbb{R}^{D}$, the parameter d in FTN is set to D, and for topic distribution $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^{K}$, it is set to K.

FTN can model not only the sentence-level interactions between argument representations but also the relevance between topic-level representations, which can be regarded as topic-level interactions. Finally, we concatenate the sentence-level interactions $T(\boldsymbol{R}_1, \boldsymbol{R}_2)$ and the topic-level relevance $T(\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ and send them to a two-layer neural network classifier, which first applies a nonlinear transformation and then computes the probabilities of each relation by a softmax layer.

3.4 Joint Learning

To simultaneously update the parameters in all components of TTN, we jointly tackle the topic modeling and the classification, and define the loss function of the overall model to combine the two effects as follows.

$$\mathcal{L} = \mathcal{L}_{STM} + \lambda \mathcal{L}_{MLP} \quad (10)$$

where $\mathcal{L}_{STM}$ represents the loss of STM and $\mathcal{L}_{MLP}$ is the cross entropy loss of the classifier. λ is the trade-off parameter controlling the balance

between the topic model and the MLP classifier. To prevent overfitting, a dropout operation is performed on the parameter vector input to the softmax layer.

4 Experimentation

4.1 Experiment Settings

Due to the small number of documents in CDTB-ZX, we evaluate our model on CDTB (Li et al., 2014) with 500 annotated newswire articles from CTB (Xue et al., 2005). CDTB contains 7310 annotated relations (implicit: 5496) which can be divided into 4 classes and 17 categories. To make full use of this corpus, we erase the existing connectives information and treat all samples as implicit discourse relation samples.

Following previous work (Kong and Zhou, 2017), we choose the same 450 documents as the training set and the remaining 50 documents as the testing set. We also evaluate TTN on the four top-level classes in CDTB, and transform all of the non-binary trees into left binary trees. Table 1 summarizes the statistics of the four CDTB relations, i.e., *Causality*, *Coordination*, *Elaboration*, and *Transition*.

Relation	Train	Test
Causality	1213	119
Coordination	4618	515
Elaboration	1465	151
Transition	205	11

Table 1: Statistics of the discourse relations in CDTB.

We use HanLP[1] as the NLP tool for word segmentation and POS tagging, and use the Keras[2] library to implement our model. We selected 10% of the samples from the training set as the development set to fine-tune the hyper-parameters, and only give their final settings due to space limitation.

The 300-dimensional pre-trained word embeddings are provided by Word2Vec (Mikolov et al., 2013), and the dimension of the POS embeddings is set to 50. The trade-off parameter λ in Equ. (10) is set to 1.0. To alleviate the data sparseness of the input BoW representations, we limit the vocabulary to the top 5000 most frequent words, i.e., $V = 5000$.

[1]https://github.com/hankcs/HanLP
[2]https://keras.io/

In STM, the number of topics is set to 256, and the number of neurons in the single-layer networks $f_h(\cdot), f_\mu(\cdot), f_\sigma(\cdot)$ are set to 512, 256 and 256, respectively. In addition, the generator g is implemented by a two-layer network with a hidden layer size of 512. In the GCN-based text encoder, the number of layers L is set to 3, and the convolution kernel size C is set to 3. In FTN, the number of tensor slices m is set to 128, and r of the tensor factorization is set to 10. The size of the nonlinear transformation layer in the MLP classifier and the droupout rate are set to 64 and 0.5, respectively.

4.2 Experimental Results

To exhibit the effectiveness of our TTN model, we selected **Bi-LSTM**, **CNN** and **GCN** (Dauphin et al., 2017) as baselines in addition to three state-of-the-art models proposed in previous works: (1) **Liu&Li** (Liu and Li, 2016): a multi-level attention model that simulates the repeated reading process by stacking multiple attention layers with external memory; (2) **Rönnqvist** (Rönnqvist et al., 2017): a Bi-LSTM model with attention mechanism that first links argument pairs by inserting special labels; and (3) **Guo** (Guo et al., 2018): a neural tensor network that encodes the arguments by Bi-LSTM and interactive attention. Among them, GCN uses the same settings as our model. Following Liu and Li (2016), the hidden size for each direction of Bi-LSTM is set to 350, the same as the dimension of the word embeddings. Following Qin et al. (2016), the convolution kernel size and the number in CNN are set to 2 and 1024, respectively. The three state-of-the-art models are reproduced following their corresponding work.

The experimental results on CDTB are illustrated in Table 2. It shows that our TTN model outperforms the other baselines in both the micro and macro F1-scores. This indicated that topic-level information is a vital evidence to reveal the relationships among arguments and justify the effectiveness of our TTN model.

Compared with the basic recurrent neural network Bi-LSTM, the CNN and GCN significantly improve the micro and macro F1-scores due to the powerful capabilities of convolution kernels to capture features. Especially, GCN is better than CNN because it can control the information flow in the convolutional network using gate units and extract hierarchical features by stacking multiple layers. In addition, Liu&Li and Guo, two state-of-

Model	Caus.	Coor.	Elab.	Tran.	Micro-F1	Macro-F1
Bi-LSTM	37.4	79.8	51.8	73.7	68.7	61.1
CNN	41.2	81.5	52.5	80.0	71.4	64.4
GCN	**46.2**	82.4	51.4	76.2	71.5	64.6
Liu&Li	42.8	81.4	54.6	**85.7**	71.1	66.2
Rönnqvist	39.2	81.6	57.1	78.3	71.1	64.3
Guo	42.4	80.1	60.0	80.0	70.7	65.8
TTN	40.6	**83.1**	**60.7**	84.2	**73.6**	**67.8**

Table 2: Performance of six baselines and TNN with F1-scores.

the-art models on English implicit discourse relation recognition, and Rönnqvist, a state-of-the-art model on Chinese, focus on extracting sentence-level features from arguments and achieve similar performance.

Our TTN model outperforms all of the baselines with large gains from 2.1 to 4.9 in the micro F1-score and significant gains from 1.6 to 6.7 in the macro F1-score. Compared with the baselines, TTN not only captures the interactive features at sentence-level, but also considers the topic-level relevance among arguments. This result shows that TTN can recognize the discourse relations at a higher level to improve the performance of Chinese implicit discourse relation recognition. Different from Liu&Li, TTN not only learns the argument representations by stacking multiple layers with residuals to simulate the repeated reading, but also models the deep semantic interactions through factored tensor networks. Different from Guo, TTN not only reduces the complexity of the tensor network using tensor factorization, but also models the sentence-level and topic-level interactions together.

5 Analysis and Discussion

5.1 Impact on Different Relations

Table 2 also compares the F1-scores on different relations. We can find that our TTN achieves the highest F1-scores in the *Elaboration* and *Coordination* relations, and it achieves a comparable performance in the *Transition* relation. However, it reduces the F1-score in the *Causality* relation by 5.6, compared with GCN.

To explain the reasons behind this, we conduct experiments on some variants of TTN with the results shown in Table 3. We choose the gated convolutional network (GCN) as the **Base** model with its parameters being set the same as our model. To analyze the contribution of the topic-level representation and the factored tensor modeling method separately, we add our simplified topic model (**STM**) and our factored tensor network (**FTN**) to the Base model, respectively.

The results shows that STM gives the latent topic distributions of arguments and there is a significant improvement (+8.6) in recognizing the *Elaboration* relation. The existence of an *Elaboration* relation between two arguments means that the content of one argument is a further explanation of the other, and these arguments usually have similar topic distributions. Hence, STM essentially provides additional topic distribution features to TNN, which help in recognizing the *Elaboration* relation. Equally, STM can also improve the performance of recognizing the *Coordination* relation because two arguments with the *Coordination* relation are equally important at the semantic level, and their contents describe different aspects of one thing or different parts of a certain behavior; hence, they are also similar at the topic level in most cases. However, this does not apply to the *Causality* relation and there is a large drop (-9.8) with the lowest F1-score among all four relations. The reason behind this may be due to the fact that the recognition of the *Causality* relation relies more on the logical connection, and arguments with the *Causality* relation are not similar at the topic level in most cases. Hence, STM, which simply introduces topical information to the Base model, does not help and even may harm the recognition. Take the following two arguments as examples:

[出口快速增长， (*Exports have grown rapidly,*)]$_{Arg1}$ [成为推动经济增长的重要力量。 (*become an important force driving economic growth.*)]$_{Arg2}$

$Arg1$ is the reason for $Arg2$, and hence the relation between them is *Causality*. However,

Model	Caus.	Coor.	Elab.	Tran.	Micro-F1	Macro-F1
Base(GCN)	46.2	82.4	51.4	76.2	71.5	64.6
+STM	36.4	82.9	60.0	73.7	73.1	64.1
+FTN	41.3	82.7	55.3	84.2	72.5	66.4

Table 3: Comparison of Base, STM and FTN on the F1-score.

Model	Caus.	Coor.	Elab.	Tran.	Micro-F1	Macro-F1
TTN	40.6	83.1	60.7	84.2	73.6	67.8
NTN(Guo)	39.6	82.1	56.2	84.2	72.6	66.4

Table 4: Comparison of TTN and NTN(Guo) on the F1-score.

from the perspective of the topic, the words in the two arguments revolve around the same topic of "economic growth". Therefore, our STM will directly infers the similar topic distribution from the words of these two arguments and interfere with the recognition of the *Causality* relation.

Our neural factored tensor networks (FTNs) are capable of modeling complex semantic interactions between two arguments using multiple Bilinear models and single layer neural network. Therefore, after the addition, a certain improvement has been achieved in recognizing most relations (except for *Causality*). Especially, it improves the F1-scores of the *Elaboration* and *Transition* relations by 3.9 and 8.0, respectively.

5.2 Impact of Tensor Factorization

To further verify the impact of tensor factorization, we compare it with Guo et al. (2018). Table 4 illustrates the results, where **NTN(Guo)** is a modified version of our TTN, which uses the NTN model proposed by Guo et al. (2018) to replace our FTN.

Since NTN(Guo) does not use the tensor factorization operation, its parameter number and computational complexity increase greatly. The parameters of factored tensor network in our model are reduced by approximately 20 times, compared with NTN(Guo). If it directly adopts our parameter settings, the model will have serious overfitting, and it will not even recognize the *Transition* relation, which is only a small proportion of the training set. Therefore, following (Guo et al., 2018), we set the tensor number to a very small value. It shows that NTN(Guo) has a performance degradation of 1.0 and 1.4 in micro and macro F1-scores, respectively, indicating that the tensor factorization operation in our model is very effective. In addition, our neural tensor network can set more tensor slices to model the complex interactions between two arguments.

5.3 Error Analysis

Table 5 illustrates the error statistics of our TTN model. It shows that 51.3% of the *Causality* samples, 33.8% of the *Elaboration* samples, and 18.2% of the *Transition* samples are incorrectly identified as *Coordination*. This indicates that the error mainly occurs when judging whether a sample is *Coordination*. This may be due to two reasons, which are that the number of *Coordination* samples accounts for more than half of the training set (61.6%) and that many argument pairs with *non-Coordination* relations are similar at both the text level and the topic level. Take the following two arguments as examples:

Model	Caus.	Coor.	Elab.	Tran.
Caus.	-	51.3%	15.1%	0%
Coor.	5.4%	-	7.8%	0%
Elab.	6.0%	33.8%	-	0%
Tran.	9.1%	18.2%	0%	-

Table 5: Percentages of misclassified samples.

[甘肃省积极实施科技兴农战略，推广增产措施 (*Gansu Province promotes various agricultural applicable technologies and production increase measures*)]$_{Arg1}$ [农业获得较好收成，全年粮食总产量达七十六点六亿公斤 (*Agriculture has achieved a good harvest, and the annual total grain output reached 7.66 billion kg*)]$_{Arg2}$

In above samples, since $Arg1$ is the reason for $Arg2$, the discourse relation between them is *Causality*. However, there is a strong sentence-level correlation between the words in $Arg1$ (e.g.,

"agricultural" and "production") and those in Arg2 (e.g., "harvests", "gain", and "output"). Moreover, these two arguments are all about agriculture. Therefore, there is a strong similarity in the topic distribution, too.

6 Conclusion

In this paper, we propose a topic tensor network TTN to recognize implicit discourse relations in Chinese with both the sentence-level and topic-level representations. In addition to using a GCN-based encoder to obtain the sentence-level argument representations, we train a STM to infer the latent topic distribution as the topic-level representations. Moreover, we feed the two pairs of representations to two FTNs, respectively, to model the sentence-level interactions and topic-level relevance among arguments. Evaluation on CTDB shows that our proposed TTN model significantly outperforms several state-of-the-art baselines in both micro and macro F1-scores. In the future work, we will focus on how to mine different representations for different discourse relation types and apply the topic information to other languages.

Acknowledgments

The authors would like to thank four anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant Nos. 61836007, 61772354 and 61773276.

References

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

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112.

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

National Institute of Child Health and Human Development. 2000. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*.

Andre Cianflone and Leila Kosseim. 2018. Attention for implicit discourse relation recognition. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 1946–1951.

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 (ICML)*, pages 160–167.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 933–941.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 511–521.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778.

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 (NIPS)*, pages 3167–3175.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Fang Kong and Guodong Zhou. 2017. A CDT-styled end-to-end Chinese discourse parser. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):26.

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

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 4848–4855.

Yancui Li, Fang Kong, and Guodong Zhou. 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2105–2114.

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 (EMNLP)*, pages 343–351.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 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 Conference on Artificial Intelligence (AAAI)*, pages 2750–2756.

Yang Liu, Jiajun Zhang, and Chengqing Zong. 2017. Memory augmented attention model for Chinese implicit discourse relation recognition. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data (CCL)*, pages 411–423.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 1727–1736.

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 (NIPS)*, pages 3111–3119.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for Chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 293–303.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *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 (ACL-IJCNLP)*, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *The 6th international conference on Language Resources and Evaluation (LREC)*, pages 2961—-2968.

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

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

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1278–1286.

Samuel Rönnqvist, Niko Schenk, and Christian Chiarcos. 2017. A recurrent neural model with attention for the recognition of Chinese implicit discourse relations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–262.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 484–495.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems (NIPS)*, pages 926–934.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 184–188.

Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–731.

Naiwen Xue, Fei Xia, Fudong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 shared task on shallow discourse parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning - Shared Task (CoNLL)*, pages 1–19.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3120–3131.

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

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1507–1514.