Variational Neural Discourse Relation Recognizer

Biao Zhang[1], **Deyi Xiong**[2]*, **Jinsong Su**[1], **Qun Liu**[3,4], **Rongrong Ji**[1], **Hong Duan**[1], **Min Zhang**[2]
Xiamen University, Xiamen, China 361005[1]
Provincial Key Laboratory for Computer Information Processing Technology
Soochow University, Suzhou, China 215006[2]
ADAPT Centre, School of Computing, Dublin City University[3]
Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences[4]
zb@stu.xmu.edu.cn, {jssu, rrji, hduan}@xmu.edu.cn
qun.liu@dcu.ie, {dyxiong, minzhang}@suda.edu.cn

Abstract

Implicit discourse relation recognition is a crucial component for automatic discourse-level analysis and nature language understanding. Previous studies exploit discriminative models that are built on either powerful manual features or deep discourse representations. In this paper, instead, we explore generative models and propose a variational neural discourse relation recognizer. We refer to this model as *VarNDRR*. VarNDRR establishes a directed probabilistic model with a latent continuous variable that generates both a discourse and the relation between the two arguments of the discourse. In order to perform efficient inference and learning, we introduce neural discourse relation models to approximate the prior and posterior distributions of the latent variable, and employ these approximated distributions to optimize a reparameterized variational lower bound. This allows VarNDRR to be trained with standard stochastic gradient methods. Experiments on the benchmark data set show that VarNDRR can achieve comparable results against state-of-the-art baselines without using any manual features.

1 Introduction

Discourse relation characterizes the internal structure and logical relation of a coherent text. Automatically identifying these relations not only plays an important role in discourse comprehension and generation, but also obtains wide applications in many other relevant natural language processing tasks, such as text summarization (Yoshida et al., 2014), conversation (Higashinaka et al., 2014), question answering (Verberne et al., 2007) and information extraction (Cimiano et al., 2005). Generally, discourse relations can be divided into two categories: explicit and implicit, which can be illustrated in the following example:

> *The company was disappointed by the ruling.* <u>*because*</u> *The obligation is totally unwarranted.* (adapted from wsj_0294)

With the discourse connective *because*, these two sentences display an explicit discourse relation CONTINGENCY which can be inferred easily. Once this discourse connective is removed, however, the discourse relation becomes implicit and difficult to be recognized. This is because almost no surface information in these two sentences can signal this relation. For successful recognition of this relation, in the contrary, we need to understand the deep semantic correlation between *disappointed* and *obligation* in the two sentences above. Although explicit discourse relation recognition (DRR) has made great progress (Miltsakaki et al., 2005; Pitler et al., 2008), implicit DRR still remains a serious challenge due to the difficulty in semantic analysis.

Conventional approaches to implicit DRR often treat the relation recognition as a classification problem, where discourse arguments and relations are regarded as the inputs and outputs respectively. Generally, these methods first generate a representation for a discourse, denoted as $\mathbf{x}$[1] (e.g., manual fea-

*Corresponding author

[1]Unless otherwise specified, all variables in the paper, e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are multivariate. But for notational convenience, we

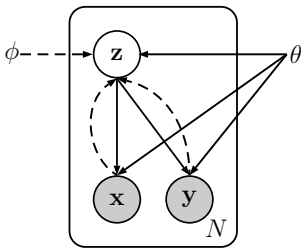


Figure 1: Graphical illustration for VarNDRR. Solid lines denote the generative model $p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z})$, dashed lines denote the variational approximation $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ to the posterior $p(\mathbf{z}|\mathbf{x},\mathbf{y})$ and $q'_\phi(\mathbf{z}|\mathbf{x})$ to the prior $p(\mathbf{z})$ for inference. The variational parameters ϕ are learned jointly with the generative model parameters θ.

tures in SVM-based recognition (Pitler et al., 2009; Lin et al., 2009) or sentence embeddings in neural networks-based recognition (Ji and Eisenstein, 2015; Zhang et al., 2015)), and then directly model the conditional probability of the corresponding discourse relation $\mathbf{y}$ given $\mathbf{x}$, i.e. $p(\mathbf{y}|\mathbf{x})$. In spite of their success, these discriminative approaches rely heavily on the goodness of discourse representation $\mathbf{x}$. Sophisticated and good representations of a discourse, however, may make models suffer from overfitting as we have no large-scale balanced data.

Instead, we assume that there is a latent continuous variable $\mathbf{z}$ from an underlying semantic space. It is this latent variable that generates both discourse arguments and the corresponding relation, i.e. $p(\mathbf{x},\mathbf{y}|\mathbf{z})$. The latent variable enables us to jointly model discourse arguments and their relations, rather than conditionally model $\mathbf{y}$ on $\mathbf{x}$. However, the incorporation of the latent variable makes the modeling difficult due to the intractable computation with respect to the posterior distribution.

Inspired by Kingma and Welling (2014) as well as Rezende et al. (2014) who introduce a variational neural inference model to the intractable posterior via optimizing a reparameterized variational lower bound, we propose a variational neural discourse relation recognizer (VarNDRR) with a latent continuous variable for implicit DRR in this paper. The key idea behind VarNDRR is that although the posterior distribution is intractable, we can approximate it via a deep neural network. Figure 1 illustrates the graph structure of VarNDRR. Specifically, there are two essential components:

- *neural discourse recognizer* As a discourse $\mathbf{x}$ and its corresponding relation $\mathbf{y}$ are independent with each other given the latent variable $\mathbf{z}$ (as shown by the solid lines), we can formulate the generation of $\mathbf{x}$ and $\mathbf{y}$ from $\mathbf{z}$ in the equation $p_\theta(\mathbf{x},\mathbf{y}|\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z})$. These two conditional probabilities on the right hand side are modeled via deep neural networks (see section 3.1).
- *neural latent approximator* VarNDRR assumes that the latent variable can be inferred from discourse arguments $\mathbf{x}$ and relations $\mathbf{y}$ (as shown by the dash lines). In order to infer the latent variable, we employ a deep neural network to approximate the posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ as well as the prior $q'_\phi(\mathbf{z}|\mathbf{x})$ (see section 3.2), which makes the inference procedure efficient. We further employ a reparameterization technique to sample z from $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ that not only bridges the gap between the recognizer and the approximator but also allows us to use the standard stochastic gradient ascent techniques for optimization (see section 3.3).

The main contributions of our work lie in two aspects. 1) We exploit a generative graphic model for implicit DRR. To the best of our knowledge, this has never been investigated before. 2) We develop a neural recognizer and two neural approximators specifically for implicit DRR, which enables both the recognition and inference to be efficient.

We conduct a series of experiments for English implicit DRR on the PDTB-style corpus to evaluate the effectiveness of our proposed VarNDRR model. Experiment results show that our variational model achieves comparable results against several strong baselines in term of F1 score. Extensive analysis on the variational lower bound further reveals that our model can indeed fit the data set with respect to discourse arguments and relations.

2 Background: Variational Autoencoder

The variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014), which forms the basis of our model, is a generative model that can be regarded as a regularized version of the standard

treat them as univariate variables in most cases. Additionally, we use bold symbols to denote variables, and plain symbols to denote values.

autoencoder. With a latent random variable $\mathbf{z}$, VAE significantly changes the autoencoder architecture to be able to capture the variations in the observed variable $\mathbf{x}$. The joint distribution of $(\mathbf{x}, \mathbf{z})$ is formulated as follows:

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z}) \tag{1}$$

where $p_\theta(\mathbf{z})$ is the prior over the latent variable, usually equipped with a simple Gaussian distribution. $p_\theta(\mathbf{x}|\mathbf{z})$ is the conditional distribution that models the probability of $\mathbf{x}$ given the latent variable $\mathbf{z}$. Typically, VAE parameterizes $p_\theta(\mathbf{x}|\mathbf{z})$ with a highly nonlinear but flexible function approximator such as a neural network.

The objective of VAE is to maximize a variational lower bound as follows:

$$\begin{aligned}\mathcal{L}_{VAE}(\theta, \phi; \mathbf{x}) = &-\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \\ &+\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \leq \log p_\theta(\mathbf{x})\end{aligned} \tag{2}$$

where $\mathrm{KL}(Q||P)$ is Kullback-Leibler divergence between two distributions Q and P. $q_\phi(\mathbf{z}|\mathbf{x})$ is an approximation of the posterior $p(\mathbf{z}|\mathbf{x})$ and usually follows a diagonal Gaussian $\mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ whose mean μ and variance σ^2 are parameterized by again, neural networks, conditioned on $\mathbf{x}$.

To optimize Eq. (2) stochastically with respect to both θ and ϕ, VAE introduces a reparameterization trick that parameterizes the latent variable $\mathbf{z}$ with the Gaussian parameters μ and σ in $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\tilde{z} = \mu + \sigma \odot \epsilon \tag{3}$$

where ϵ is a standard Gaussian variable, and $\odot$ denotes an element-wise product. Intuitively, VAE learns the representation of the latent variable not as single points, but as soft ellipsoidal regions in latent space, forcing the representation to fill the space rather than memorizing the training data as isolated representations. With this trick, the VAE model can be trained through standard backpropagation technique with stochastic gradient ascent.

3 The VarNDRR Model

This section introduces our proposed VarNDRR model. Formally, in VarNDRR, there are two observed variables, $\mathbf{x}$ for a discourse and $\mathbf{y}$ for the corresponding relation, and one latent variable $\mathbf{z}$. As illustrated in Figure 1, the joint distribution of the three variables is formulated as follows:

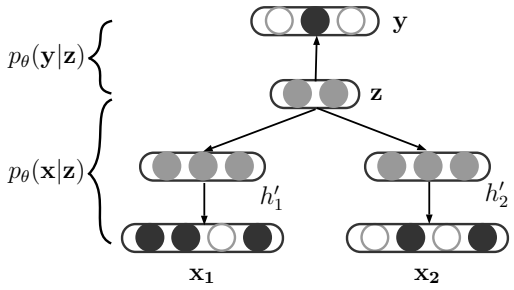


Figure 2: Neural networks for conditional probabilities $p_\theta(\mathbf{x}|\mathbf{z})$ and $p_\theta(\mathbf{y}|\mathbf{z})$. The gray color denotes real-valued representations while the white and black color 0-1 representations.

$$p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{x}, \mathbf{y}|\mathbf{z})p(\mathbf{z}) \tag{4}$$

We begin with this distribution to elaborate the major components of VarNDRR.

3.1 Neural Discourse Recognizer

The conditional distribution $p(\mathbf{x}, \mathbf{y}|\mathbf{z})$ in Eq. (4) shows that both discourse arguments and the corresponding relation are generated from the latent variable. As shown in Figure 1, $\mathbf{x}$ is d-separated from $\mathbf{y}$ by $\mathbf{z}$. Therefore the discourse $\mathbf{x}$ and the corresponding relation $\mathbf{y}$ is independent given the latent variable $\mathbf{z}$. The joint probability can be therefore formulated as follows:

$$p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \tag{5}$$

We use a neural model $q'_\phi(\mathbf{z}|\mathbf{x})$ to approximate the prior $p(\mathbf{z})$ conditioned on the discourse $\mathbf{x}$ (see the following section). With respect to the other two conditional distributions, we parameterize them via neural networks as shown in Figure 2.

Before we describe these neural networks, it is necessary to briefly introduce how discourse relations are annotated in our training data. The PDTB corpus, used as our training data, annotates implicit discourse relations between two neighboring arguments, namely *Arg1* and *Arg2*. In VarNDRR, we represent the two arguments with bag-of-word representations, and denote them as $\mathbf{x_1}$ and $\mathbf{x_2}$.

To model $p_\theta(\mathbf{x}|\mathbf{z})$ (the bottom part in Figure 2), we project the representation of the latent variable

$z \in \mathbb{R}^{d_z}$ onto a hidden layer:

$$\begin{aligned} h'_1 &= f(W_{h'_1} z + b_{h'_1}) \\ h'_2 &= f(W_{h'_2} z + b_{h'_1}) \end{aligned} \tag{6}$$

where $h'_1 \in \mathbb{R}^{d_{h'_1}}, h'_2 \in \mathbb{R}^{d_{h'_2}}$, W_* is the transformation matrix, b_* is the bias term, d_u denotes the dimensionality of vector representations of u and $f(\cdot)$ is an element-wise activation function, such as $tanh(\cdot)$, which is used throughout our model.

Upon this hidden layer, we further stack a Sigmoid layer to predict the probabilities of corresponding discourse arguments:

$$\begin{aligned} x'_1 &= \text{Sigmoid}(W_{x'_1} h'_1 + b_{x'_1}) \\ x'_2 &= \text{Sigmoid}(W_{x'_2} h'_2 + b_{x'_2}) \end{aligned} \tag{7}$$

here, $x'_1 \in \mathbb{R}^{d_{x_1}}$ and $x'_2 \in \mathbb{R}^{d_{x_2}}$ are the real-valued representations of the reconstructed x_1 and x_2 respectively.[2] We assume that $p_\theta(\mathbf{x}|\mathbf{z})$ is a multivariate Bernoulli distribution because of the bag-of-word representation. Therefore the logarithm of $p(x|z)$ is calculated as the sum of probabilities of words in discourse arguments as follows:

$$\begin{aligned} &\log\ p(x|z) \\ &= \sum_i x_{1,i} \log x'_{1,i} + (1 - x_{1,i}) \log(1 - x'_{1,i}) \\ &+ \sum_j x_{2,j} \log x'_{2,j} + (1 - x_{2,j}) \log(1 - x'_{2,j}) \end{aligned} \tag{8}$$

where $u_{i,j}$ is the jth element in u_i.

In order to estimate $p_\theta(\mathbf{y}|\mathbf{z})$ (the top part in Figure 2), we stack a softmax layer over the multilayer-perceptron-transformed representation of the latent variable z:

$$y' = \text{SoftMax}(W_{y'} \text{MLP}(z) + b_{y'}) \tag{9}$$

$y' \in \mathbb{R}^{d_y}$, and d_y denotes the number of discourse relations. MLP projects the representation of latent variable $\mathbf{z}$ into a d_m-dimensional space through four internal layers, each of which has dimension d_m. Suppose that the true relation is $y \in \mathbb{R}^{d_y}$, the logarithm of $p(y|z)$ is defined as:

$$\log\ p(y|z) = \sum_{i=1}^{d_y} y_i \log y'_i \tag{10}$$

[2]Notice that the equality of $d_{x_1} = d_{x_2}, d_{h'_1} = d_{h'_2}$ is not necessary though we assume so in our experiments.

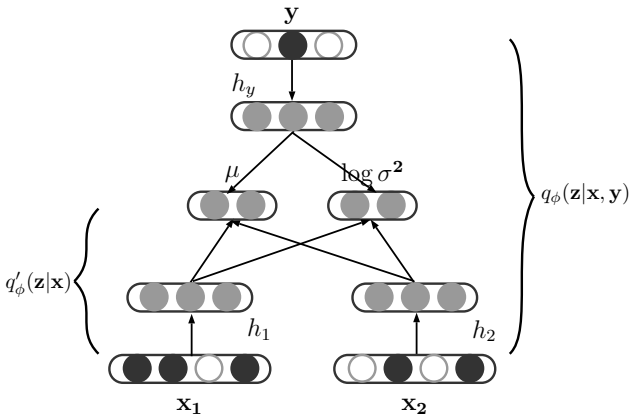


Figure 3: Neural networks for Gaussian parameters μ and $\log \sigma$ in the approximated posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ and prior $q'_\phi(\mathbf{z}|\mathbf{x})$.

In order to precisely estimate these conditional probabilities, our model will force the representation z of the latent variable to encode semantic information for both the reconstructed discourse x' (Eq. (8)) and predicted discourse relation y' (Eq. (10)), which is exactly what we want.

3.2 Neural Latent Approximator

For the joint distribution in Eq. (5), we can define a variational lower bound that is similar to Eq. (2). The difference lies in two aspects: the approximate prior $q'_\phi(\mathbf{z}|\mathbf{x})$ and posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$. We model both distributions as a multivariate Gaussian distribution with a diagonal covariance structure:

$$\mathcal{N}(\mathbf{z}; \mu, \sigma^2 \mathbf{I})$$

The mean μ and s.d. σ of the approximate distribution are the outputs of the neural network as shown in Figure 3, where the prior and posterior have different conditions and independent parameters.

Approximate Posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})$ is modeled to condition on both observed variables: the discourse arguments $\mathbf{x}$ and relations $\mathbf{y}$. Similar to the calculation of $p_\theta(\mathbf{x}|\mathbf{z})$, we first transform the input $\mathbf{x}$ and $\mathbf{y}$ into a hidden representation:

$$\begin{aligned} h_1 &= f(W_{h_1} x_1 + b_{h_1}) \\ h_2 &= f(W_{h_2} x_2 + b_{h_2}) \\ h_y &= f(W_{h_y} y + b_{h_y}) \end{aligned} \tag{11}$$

where $h_1 \in \mathbb{R}^{d_{h_1}}, h_2 \in \mathbb{R}^{d_{h_2}}$ and $h_y \in \mathbb{R}^{d_{h_y}}$.[3]

[3]Notice that d_{h_1}/d_{h_2} are not necessarily equal to $d_{h'_1}/d_{h'_2}$.

We then obtain the Gaussian parameters of the posterior μ and $\log \sigma^2$ through linear regression:

$$\begin{aligned} \mu &= W_{\mu_1} h_1 + W_{\mu_2} h_2 + W_{\mu_y} h_y + b_\mu \\ \log \sigma^2 &= W_{\sigma_1} h_1 + W_{\sigma_2} h_2 + W_{\sigma_y} h_y + b_\sigma \end{aligned} \quad (12)$$

where $\mu, \sigma \in \mathbb{R}^{d_z}$. In this way, this posterior approximator can be efficiently computed.

Approximate Prior $q'_\phi(\mathbf{z}|\mathbf{x})$ is modeled to condition on discourse arguments $\mathbf{x}$ alone. This is based on the observation that discriminative models are able to obtain promising results using only $\mathbf{x}$. Therefore, assuming the discourse arguments encode the prior information for discourse relation recognition is meaningful.

The neural model for prior $q'_\phi(\mathbf{z}|\mathbf{x})$ is the same as that (i.e. Eq (11) and (12)) for posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$ (see Figure 3), except for the absence of discourse relation $\mathbf{y}$. For clarity , we use μ' and σ' to denote the mean and s.d. of the approximate prior.

With the parameters of Gaussian distribution, we can access the representation z using different sampling strategies. However, traditional sampling approaches often breaks off the connection between recognizer and approximator, making the optimization difficult. Instead, we employ the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014) as in Eq. (3). During training, we sample the latent variable using $\tilde{z} = \mu + \sigma \odot \epsilon$; during testing, however, we employ the expectation of $\mathbf{z}$ in the approximate prior distribution, i.e. set $\tilde{z} = \mu'$ to avoid uncertainty.

3.3 Parameter Learning

We employ the Monte Carlo method to estimate the expectation over the approximate posterior, that is $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{x}, \mathbf{y}|\mathbf{z})]$. Given a training instance $(x^{(t)}, y^{(t)})$, the joint training objective is defined:

$$\begin{aligned} \mathcal{L}(\theta, \phi) \simeq & -\mathrm{KL}(q_\phi(\mathbf{z}|x^{(t)}, y^{(t)})||q'_\phi(\mathbf{z}|x^{(t)})) \\ & + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x^{(t)}, y^{(t)}|\tilde{z}^{(t,l)}) \end{aligned} \quad (13)$$

where $\tilde{z}^{(t,l)} = \mu^{(t)} + \sigma^{(t)} \odot \epsilon^{(l)}$ and $\epsilon^{(l)} \sim \mathcal{N}(0, \mathbf{I})$

L is the number of samples. The first term is the KL divergence of two Gaussian distributions which can be computed and differentiated without estimation.

Algorithm 1 Parameter Learning Algorithm of VarNDRR.

Inputs: A, the maximum number of iterations;
M, the number of instances in one batch;
L, the number of samples;
$\theta, \phi \leftarrow$ Initialize parameters
repeat
  $\mathcal{D} \leftarrow$ getRandomMiniBatch(M)
  $\epsilon \leftarrow$ getRandomNoiseFromStandardGaussian()
  $g \leftarrow \nabla_{\theta,\phi} \mathcal{L}(\theta, \phi; \mathcal{D}, \epsilon)$
  $\theta, \phi \leftarrow$ parameterUpdater($\theta, \phi; g$)
until convergence of parameters (θ, ϕ) or reach the maximum iteration A

Relation	#Instance Number		
	Train	Dev	Test
COM	1942	197	152
CON	3342	295	279
EXP	7004	671	574
TEM	760	64	85

Table 1: Statistics of implicit discourse relations for the training (Train), development (Dev) and test (Test) sets in PDTB.

Maximizing this objective will minimize the difference between the approximate posterior and prior, thus making the setting $\tilde{z} = \mu'$ during testing reasonable. The second term is the approximate expectation of $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}[\log p_\theta(\mathbf{x}, \mathbf{y}|\mathbf{z})]$, which is also differentiable.

As the objective function in Eq. (13) is differentiable, we can optimize both the model parameters θ and variational parameters ϕ jointly using standard gradient ascent techniques. The training procedure for VarNDRR is summarized in Algorithm 1.

4 Experiments

We conducted experiments on English implicit DRR task to validate the effectiveness of VarNDRR.[4]

4.1 Dataset

We used the largest hand-annotated discourse corpus *PDTB 2.0*[5] (Prasad et al., 2008) (PDTB hereafter). This corpus contains discourse annotations over 2,312 Wall Street Journal articles, and is organized in different sections. Following previous work (Pitler et al., 2009; Zhou et al., 2010; Lan et

[4]Source code is available at https://github.com/DeepLearnXMU/VarNDRR.

[5]http://www.seas.upenn.edu/ pdtb/

Model	Acc	P	R	F1
R & X (2015)	-	-	-	41.00
J & E (2015)	70.27	-	-	35.93
SVM	63.10	22.79	64.47	33.68
SCNN	60.42	22.00	67.76	33.22
VarNDRR	63.30	24.00	71.05	35.88

(a) COM vs Other

Model	Acc	P	R	F1
(R & X (2015))	-	-	-	53.80
(J & E (2015))	76.95	-	-	52.78
SVM	62.62	39.14	72.40	50.82
SCNN	63.00	39.80	75.29	52.04
VarNDRR	53.82	35.39	88.53	50.56

(b) CON vs Other

Model	Acc	P	R	F1
(R & X (2015))	-	-	-	69.40
(J & E (2015))	69.80	-	-	80.02
SVM	60.71	65.89	58.89	62.19
SCNN	63.00	56.29	91.11	69.59
VarNDRR	57.36	56.46	97.39	71.48

(c) EXP vs Other

Model	Acc	P	R	F1
(R & X (2015))	-	-	-	33.30
(J & E (2015))	87.11	-	-	27.63
SVM	66.25	15.10	68.24	24.73
SCNN	76.95	20.22	62.35	30.54
VarNDRR	62.14	17.40	97.65	29.54

(d) TEM vs Other

Table 2: Classification results of different models on the implicit DRR task. **Acc**=Accuracy, **P**=Precision, **R**=Recall, and **F1**=F1 score.

al., 2013; Zhang et al., 2015), we used sections 2-20 as our training set, sections 21-22 as the test set. Sections 0-1 were used as the development set for hyperparameter optimization.

In PDTB, discourse relations are annotated in a predicate-argument view. Each discourse connective is treated as a predicate that takes two text spans as its arguments. The discourse relation tags in PDTB are arranged in a three-level hierarchy, where the top level consists of four major semantic *classes*: TEMPORAL (TEM), CONTINGENCY (CON), EXPANSION (EXP) and COMPARISON (COM). Because the top-level relations are general enough to be annotated with a high inter-annotator agreement and are common to most theories of discourse, in our experiments we only use this level of annotations.

We formulated the task as four separate one-against-all binary classification problems: each top level class vs. the other three discourse relation classes. We also balanced the training set by resampling training instances in each class until the number of positive and negative instances are equal. In contrast, all instances in the test and development set are kept in nature. The statistics of various data sets is listed in Table 1.

4.2 Setup

We tokenized all datasets using *Stanford NLP Toolkit*[6]. For optimization, we employed the Adam algorithm (Kingma and Ba, 2014) to update parameters. With respect to the hyperparameters M, L, A and the dimensionality of all vector representations, we set them according to previous work (Kingma and Welling, 2014; Rezende et al., 2014) and preliminary experiments on the development set. Finally, we set $M = 16, A = 1000, L = 1, d_z = 20, d_{x_1} = d_{x_2} = 10001, d_{h_1} = d_{h_2} = d_{h'_1} = d_{h'_2} = d_m = d_{h_y} = 400, d_y = 2$ for all experiments.[7]. All parameters of VarNDRR are initialized by a Gaussian distribution ($\mu = 0, \sigma = 0.01$). For Adam, we set $\beta_1 = 0.9$, $\beta_2 = 0.999$ with a learning rate 0.001. Additionally, we tied the following parameters in practice: W_{h_1} and W_{h_2}, $W_{x'_1}$ and $W_{x'_2}$.

We compared VarNDRR against the following two different baseline methods:

- **SVM:** a support vector machine (SVM) classifier[8] trained with several manual features.
- **SCNN:** a shallow convolutional neural network proposed by Zhang et al. (2015).

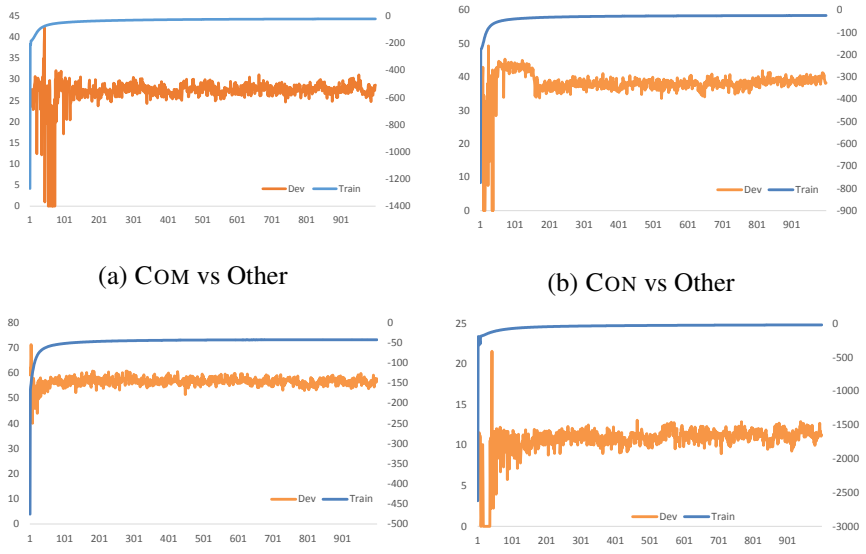
We also provide results from two state-of-the-art systems:

- **Rutherford and Xue (2015)** convert explicit discourse relations into implicit instances.
- **Ji and Eisenstein (2015)** augment discourse representations via entity connections.

[6]http://nlp.stanford.edu/software/corenlp.shtml

[7]There is one dimension in d_{x_1} and d_{x_2} for unknown words.

[8]http://svmlight.joachims.org/



(a) COM vs Other

(b) CON vs Other

(c) EXP vs Other

(d) TEM vs Other

Figure 4: Illustration of the variational lower bound (blue color) on the training set and F-score (brown color) on the development set. Horizontal axis: the epoch numbers; Vertical axis: the F1 score for relation classification (left) and the estimated average variational lower bound per datapoint (right).

Features used in **SVM** are taken from the state-of-the-art implicit discourse relation recognition model, including *Bag of Words*, *Cross-Argument Word Pairs*, *Polarity*, *First-Last, First3*, *Production Rules*, *Dependency Rules* and *Brown cluster pair* (Rutherford and Xue, 2014). In order to collect bag of words, production rules, dependency rules, and cross-argument word pairs, we used a frequency cut-off of 5 to remove rare features, following Lin et al. (2009).

4.3 Classification Results

Because the development and test sets are imbalanced in terms of the ratio of positive and negative instances, we chose the widely-used F1 score as our major evaluation metric. In addition, we also provide the precision, recall and accuracy for further analysis. Table 2 summarizes the classification results.

From Table 2, we observe that the proposed VarNDRR outperforms **SVM** on COM/EXP/TEM and **SCNN** on EXP/COM according to their F1 scores. Although it fails on CON, VarNDRR achieves the best result on EXP/COM among these three models. Overall, VarNDRR is competitive in comparison with these two baselines. With respect to the accuracy, our model does not yield substantial improvements over the two baselines. This may be because that we used the F1 score rather than the accuracy, as our selection criterion on the development set. With respect to the precision and recall, our model tends to produce relatively lower precisions but higher recalls. This suggests that the improvements of VarNDRR in terms of F1 scores mostly benefits from the recall values.

Comparing with the state-of-the-art results of previous work (Ji and Eisenstein, 2015; Rutherford and Xue, 2015), VarNDRR achieves comparable results in term of the F1 scores. Specifically, VarNDRR outperforms Rutherford and Xue (2015) on EXP, and Ji and Eisenstein (2015) on TEM. However, the accuracy of our model fails to surpass these models. We argue that this is because both baselines use many manual features designed with prior human knowledge, but our model is purely neural-based.

Additionally, we find that the performance of our model is proportional to the number of training instances. This suggests that collecting more training instances (in spite of the noises) may be beneficial to our model.

4.4 Variational Lower Bound Analysis

In addition to the classification performance, the efficiency in learning and inference is another concern

for variational methods. Figure 4 shows the training procedure for four tasks in terms of the variational lower bound on the training set. We also provide F1 scores on the development set to investigate the relations between the variational lower bound and recognition performance.

We find that our model converges toward the variational lower bound considerably fast in all experiments (within 100 epochs), which resonates with the previous findings (Kingma and Welling, 2014; Rezende et al., 2014). However, the change trend of the F1 score does not follow that of the lower bound which takes more time to converge. Particularly to the four discourse relations, we further observe that the change paths of the F1 score are completely different. This may suggest that the four discourse relations have different properties and distributions.

In particular, the number of epochs when the best F1 score reaches is also different for the four discourse relations. This indicates that dividing the implicit DRR into four different tasks according to the type of discourse relations is reasonable and better than performing DRR on the mixtures of the four relations.

5 Related Work

There are two lines of research related to our work: *implicit discourse relation recognition* and *variational neural model*, which we describe in succession.

Implicit Discourse Relation Recognition Due to the release of Penn Discourse Treebank (Prasad et al., 2008) corpus, constantly increasing efforts are made for implicit DRR. Upon this corpus, Pilter et al. (2009) exploit several linguistically informed features, such as polarity tags, modality and lexical features. Lin et al. (2009) further incorporate context words, word pairs as well as discourse parse information into their classifier. Following this direction, several more powerful features have been exploited: entities (Louis et al., 2010), word embeddings (Braud and Denis, 2015), Brown cluster pairs and co-reference patterns (Rutherford and Xue, 2014). With these features, Park and Cardie (2012) perform feature set optimization for better feature combination.

Different from feature engineering, predicting discourse connectives can indirectly help the relation classification (Zhou et al., 2010; Patterson and Kehler, 2013). In addition, selecting explicit discourse instances that are similar to the implicit ones can enrich the training corpus for implicit DRR and gains improvement (Wang et al., 2012; Lan et al., 2013; Braud and Denis, 2014; Fisher and Simmons, 2015; Rutherford and Xue, 2015). Very recently, neural network models have been also used for implicit DRR due to its capability for representation learning (Ji and Eisenstein, 2015; Zhang et al., 2015).

Despite their successes, most of them focus on the discriminative models, leaving the field of generative models for implicit DRR a relatively uninvestigated area. In this respect, the most related work to ours is the latent variable recurrent neural network recently proposed by Ji et al. (2016). However, our work differs from theirs significantly, which can be summarized in the following three aspects: 1) they employ the recurrent neural network to represent the discourse arguments, while we use the simple feed-forward neural network; 2) they treat the discourse relations directly as latent variables, rather than the underlying semantic representation of discourses; 3) their model is optimized in terms of the data likelihood, since the discourse relations are observed during training. However, VarNDRR is optimized under the variational theory.

Variational Neural Model In the presence of continuous latent variables with intractable posterior distributions, efficient inference and learning in directed probabilistic models is required. Kingma and Welling (2014) as well as Rezende et al. (2014) introduce variational neural networks that employ an approximate inference model for intractable posterior and reparameterized variational lower bound for stochastic gradient optimization. Kingma et al. (2014) revisit the approach to semi-supervised learning with generative models and further develop new models that allow effective generalization from a small labeled dataset to a large unlabeled dataset. Chung et al. (2015) incorporate latent variables into the hidden state of a recurrent neural network, while Gregor et al. (2015) combine a novel spatial attention mechanism that mimics the foveation of human eyes, with a sequential variational auto-encoding framework that allows the iterative construction of

complex images.

We follow the spirit of these variational models, but focus on the adaptation and utilization of them onto implicit DRR, which, to the best of our knowledge, is the first attempt in this respect.

6 Conclusion and Future Work

In this paper, we have presented a variational neural discourse relation recognizer for implicit DRR. Different from conventional discriminative models that directly calculate the conditional probability of the relation $\mathbf{y}$ given discourse arguments $\mathbf{x}$, our model assumes that it is a latent variable from an underlying semantic space that generates both $\mathbf{x}$ and $\mathbf{y}$. In order to make the inference and learning efficient, we introduce a neural discourse recognizer and two neural latent approximators as our generative and inference model respectively. Using the reparameterization technique, we are able to optimize the whole model via standard stochastic gradient ascent algorithm. Experiment results in terms of classification and variational lower bound verify the effectiveness of our model.

In the future, we would like to exploit the utilization of discourse instances with explicit relations for implicit DRR. For this we can start from two directions: 1) converting explicit instances into pseudo implicit instances and retraining our model; 2) developing a semi-supervised model to leverage semantic information inside discourse arguments. Furthermore, we are also interested in adapting our model to other similar tasks, such as nature language inference.

Acknowledgments

The authors were supported by National Natural Science Foundation of China (Grant Nos 61303082, 61672440, 61402388, 61622209 and 61403269), Natural Science Foundation of Fujian Province (Grant No. 2016J05161), Natural Science Foundation of Jiangsu Province (Grant No. BK20140355), Research fund of the Provincial Key Laboratory for Computer Information Processing Technology in Soochow University (Grant No. KJS1520), and Research fund of the Key Laboratory for Intelligence Information Processing in the Institute of Computing Technology of the Chinese Academy of Sciences (Grant No. IIP2015-4). We also thank the anonymous reviewers for their insightful comments.

References

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proc. of COLING*, pages 1694–1705, August.

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

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Proc. of NIPS*.

Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. 2005. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering*, 55:59–83.

Robert Fisher and Reid Simmons. 2015. Spectral semi-supervised discourse relation classification. In *Proc. of ACL-IJCNLP*, pages 89–93, July.

Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. *CoRR*, abs/1502.04623.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proc. of COLING*, pages 928–939.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *TACL*, pages 329–344.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proc. of NAACL*, pages 332–342, June.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proc. of ICLR*.

Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Proc. of NIPS*, pages 3581–3589.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition. In *Proc. of ACL*, pages 476–485, Sofia, Bulgaria, August.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proc. of EMNLP*, pages 343–351.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proc. of SIGDIAL*, pages 59–62, Tokyo, Japan, September.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proc. of TLT2005*.

Joonsuk Park and Claire Cardie. 2012. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. In *Proc. of SIGDIAL*, pages 108–112, Seoul, South Korea, July.

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *Proc. of EMNLP*, pages 914–923.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

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

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 *LREC*. Citeseer.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. of ICML*, pages 1278–1286.

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

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proc. of NAACL-HLT*, pages 799–808, May–June.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proc. of SIGIR*, pages 735–736.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. 2012. Implicit discourse relation recognition by selecting typical training examples. In *Proc. of COLING*, pages 2757–2772.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proc. of EMNLP*, pages 1834–1839, October.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proc. of EMNLP*, September.

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 *Proc. of COLING*, pages 1507–1514.