Learning Event Graph Knowledge for Abductive Reasoning

Li Du, Xiao Ding*, Ting Liu, and Bing Qin
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{ldu, xding, tliu, qinb}@ir.hit.edu.cn

Abstract

Abductive reasoning aims at inferring the most plausible explanation for observed events, which would play critical roles in various NLP applications, such as reading comprehension and question answering. To facilitate this task, a narrative text based abductive reasoning task αNLI is proposed, together with explorations about building reasoning framework using pretrained language models. However, abundant event commonsense knowledge is not well exploited for this task. To fill this gap, we propose a variational autoencoder based model ege-RoBERTa, which employs a latent variable to capture the necessary commonsense knowledge from event graph for guiding the abductive reasoning task. Experimental results show that through learning the external event graph knowledge, our approach outperforms the baseline methods on the αNLI task.

1 Introduction

Abductive reasoning aims at seeking for the best explanations for incomplete observations (Bhagavatula et al., 2019). For example, given observations *Forgot to close window when leaving home* and *The room was in a mess*, human beings can generate a reasonable hypothesis for explaining the observations, such as *A thief entered the room* based on commonsense knowledge in their mind. However, due to the lack of commonsense knowledge and effective reasoning mechanism, this is still a challenging problem for today's cognitive intelligent systems (Charniak and Shimony, 1990; Oh et al., 2013; Kruengkrai et al., 2017).

Most previous works focus on conducting abductive reasoning based on formal logic (Eshghi et al., 1988; Levesque, 1989; Ng et al., 1990; Paul, 1993). However, the rigidity of formal logic limits the application of abductive reasoning in NLP tasks, as it is hard to express the complex semantics of natural language in a formal logic system. To facilitate this, Bhagavatula et al. (2019) proposed a natural language based abductive reasoning task αNLI. As shown in Figure 1 (a), given two observed events O_1 and O_2, the αNLI task requires the prediction model to choose a more reasonable explanation from two candidate hypothesis events H_1 and H_2. Both observed events and hypothesis events are daily-life events, and are described in natural language. Together with the αNLI task, Bhagavatula et al. (2019) also explored conducting such reasoning using pretrained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019).

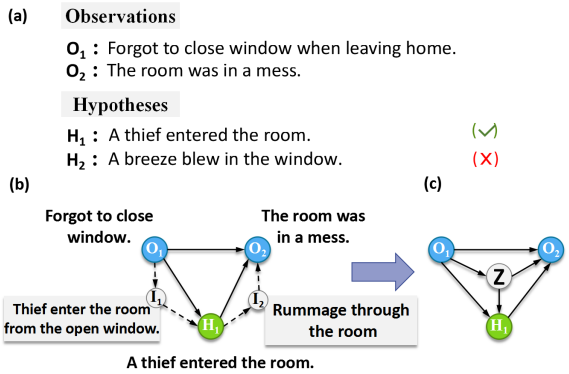

Figure 1: (a) An example of abductive reasoning. (b) Additional commonsense knowledge (such as event I_1 and I_2) is necessary for inferring the correct hypothesis. Such knowledge could be described using an event graph. (c) A latent variable z is employed to learn the commonsense knowledge from event graph.

However, despite pretrained language models could capture rich linguistic knowledge benefit for understanding the semantics of events, additional commonsense knowledge is still necessary for the abductive reasoning. For example, as illustrated in Figure 1 (b), given observations O_1 and O_2, to choose the more likely explanation H_1 : *A thief*

*Corresponding author

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5181–5190
August 1–6, 2021.

entered the room and exclude H_2 : *A breeze blew in the window*, prediction model should have the commonsense knowledge that it is hardly possible for a breeze to mess up the room, whereas *a thief may enter the room from the open window* (I_1), then *rummage through the room* (I_2) and lead to a mess. These intermediary events (I_1 and I_2) can serve as necessary commonsense knowledge for understanding the relationship between observed events and hypothesis events.

We notice that the observed events, hypothesis events, intermediary events and their relationships could be described using an event graph, which can be constructed based on an auxiliary dataset. The challenge is how to learn such commonsense knowledge from the constructed event graph.

To address this issue, we propose an Event Graph Enhanced RoBERTa (ege-RoBERTa) model, and a two-stage training procedure. Specifically, as shown in Figure 1 (c), on the basis of the RoBERTa framework, we additionally introduce a latent variable z to model the information about the intermediary events. In the pretraining stage, ege-RoBERTa is trained upon an event-graph-based pseudo instance set to capture the commonsense knowledge using the latent variable z. In the finetuning stage, model adapts the commonsense knowledge captured by z to conduct the abductive reasoning.

Experimental results show that ege-RoBERTa could effectively learn the commonsense knowledge from a well-designed event graph, and improve the model performance on the αNLI task compared to the baseline methods. The code is released at https://github.com/sjcfr/ege-RoBERTa.

2 Background

2.1 Problem Formalization

As shown in Figure 1 (a), αNLI can be defined as a multiple-choice task. Given two observed events O_1 and O_2 happened in a sequential order, one needs to choose a more reasonable hypothesis event from two candidates H_1 and H_2 for explaining the observations. Therefore, we formalize the abductive reasoning task as a conditional distribution $p(Y|O_1, H_i, O_2)$, where $H_i \in \{H_1, H_2\}$, and $Y \in [0, 1]$ is a relatedness score measuring the reasonableness of H_i.

In the αNLI dataset, H_i is set to be an explanation event happens intermediate to O_1 and O_2 (Bhagavatula et al., 2019). Hence, O_1, O_2 and H_i form an event temporal sequence O_1, H_i, O_2. For brevity, we denote the event sequence as $X = (O_1, H_i, O_2)$. Therefore, taking the event order into consideration, we further characterize the abductive reasoning task as $p(Y|X)$.

2.2 Event Graph

Formally, an event graph could be denoted as $G = \{V, R\}$, where V is the node set, and R is the edge set. Each node $V_i \in V$ corresponds to an event, while $R_{ij} \in R$ is a directed edge $V_i \rightarrow V_j$ along with a weight W_{ij}, which denotes the probability that V_j is the subsequent event of V_i.

Given observed events and a certain hypothesis event, from the event graph we could acquire additional commonsense knowledge about: **(1) the intermediary events, (2) the relationships between events**. As Figure 1 (b) shows, the observed events, hypothesis event and intermediary events compose another event sequence $(O_1, I_1, H_i, I_2, O_2)$. For clarity, we define such event sequence as posterior event sequence X', where $X' = (O_1, I_1, H_i, I_2, O_2)$. The relationship between events within X' could be described by an adjacency matrix $A \in \mathbb{R}^{5\times5}$, with each element initialized using the edge weights of the event graph:

$$A_{jk} = \begin{cases} W_{jk}, & \text{if } V_j \rightarrow V_k \in R, \\ 0, & \text{others.} \end{cases} \quad (1)$$

The matrix A could describe the adjacency relationship between arbitrary two events in X'.

3 Ege-RoBERTa as a Conditional Variational Autoencoder Based Reasoning Framework

In this paper, rather than directly predicts the relatedness score Y based on the event sequence X, we propose to predict Y based on both X and additional commonsense knowledge (i.e. posterior event sequence X' and adjacency matrix A). To this end, we introduce a latent variable z to learn such knowledge from an event graph through a two stage training procedure. To effectively capture the event graph knowledge through z and conduct the abductive reasoning task based on z, we frame the ege-RoBERTa model as a conditional variational autoencoder (CVAE) (Sohn et al., 2015).

Specifically, with regard to the latent variable z, ege-RoBERTa characterizes the conditional distribution $P(Y|X)$ using three neural networks: a prior network $p_\theta(z|X)$, a recognition network $q_\phi(z|X', A)$ and a neural likelihood $p_\theta(Y|X, z)$,

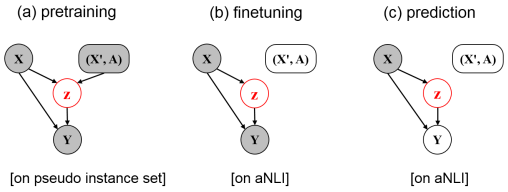

Figure 2: Illustration of the pretraining, finetuning and prediction process of ege-RoBERTa. The grey color in circle denotes the availability of corresponding information. For example, in the pretraining stage conducted on the pseudo instance set, X, Y and additional commonsense knowledge X' and A are available. While in the finetuning stage on αNLI, X' and A are absent.

where θ and ϕ denote the parameters of networks. Moreover, instead of directly maximize $P(Y|X)$, following CVAE (Sohn et al., 2015), ege-RoBERTa proposes to maximize the evidence lower bound (ELBO) of $P(Y|X)$:

$$\begin{aligned} L^{ELBO}(\theta,\phi) =& \mathbb{E}_{q_\phi(z|X',A)}\log(p_\theta(Y|X,z)) \\ & - \mathrm{KL}(q_\phi(z|X',A)||p_\theta(z|X)) \\ \leq & \log p(Y|X) \end{aligned} \quad (2)$$

Note that, in the recognition network, the latent variable z is directly conditioned on X' and A, where $X' = \{O_1, I_1, H_i, I_2, O_2\}$ is the posterior event sequence, A is an adjacency matrix describing the relationship between events within X'. This enables z to capture the event graph knowledge from X' and A. Through minimizing the KL term of ELBO, we can teach the prior network $p_\theta(z|X)$ to learn the event graph knowledge from the recognition network as much as possible. Then in the neural likelihood $p_\theta(Y|X,z)$ the relatedness score Y could be predicted based on X and z, which captures the event graph knowledge.

However, the event graph knowledge is absent in the αNLI dataset. To learn such knowledge, we design the following two-stage training procedure:

Pre-training Stage: Learning Event Graph Knowledge from a Pseudo Instance Set In this stage, ege-RoBERTa is pretrained on a prebuilt event-graph-based pseudo instance set, which contains rich information about the intermediary events and the events relationships. As shown in Figure 2 (a), the latent variable z is directly conditioned on X' and A. Therefore, z could be employed to learn the event graph knowledge.

Finetuning Stage: Adapt Event Graph Knowledge to the Abductive Reasoning Task As Figure 2 (b) shows, at the finetuning stage, ege-RoBERTa is trained on the αNLI dataset without the additional information X' and A. In this stage model learns to adapt the captured event graph knowledge to the abductive reasoning task. Then as Figure 2 (c) shows, after the two-stage training process, ege-RoBERTa could predict the relatedness score Y based on the latent variable z.

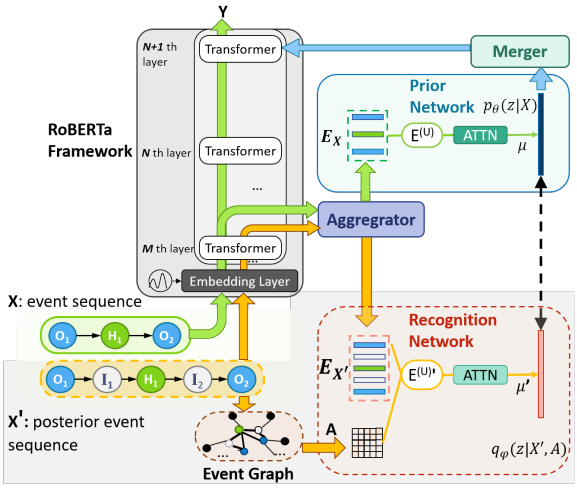

Figure 3: Architecture of ege-RoBERTa.

4 Architecture of ege-RoBERTa

We introduce the specific implementation of ege-RoBERTa. As illustrated in Figure 3, ege-RoBERTa introduces four modules in addition to the RoBERTa framework: (1) an aggregator providing representation for any event within X and X'; (2) an attention-based prior network for modeling $p_\theta(z|X)$; (3) a graph neural network based recognition network for modeling $q_\phi(z|X',A)$; (4) a merger to merge the latent variable z into RoBERTa frame for downstream abductive reasoning task.

4.1 Event Representation Aggregator

The event representation aggregator provides distributed representation for events in both the event sequence X and the posterior event sequence X'. To this end, the aggregator employs attention mechanism to aggregate token representations of the event sequence from hidden states of RoBERTa.

Given an event sequence X composed of tokens $[[\text{CLS}], (x_1^1,\ldots,x_{l_1}^1),\ldots,(x_1^3,\ldots,x_{l_3}^3)]$ (where [CLS] is the special classification token (Devlin et al., 2019), and x_k^j is the kth token within the jth event), the Mth transformer layer of RoBERTa encodes these tokens into contextualized distributed representations $H^{(M)} = [h_{[CLS]}, (h_1^1,\ldots,h_{l_1}^1),\ldots,(h_1^3,\ldots,h_{l_3}^3)]$, where $h_k^j \in \mathbb{R}^{1\times d}$ is the distributed representation of the kth token within the jth event. Then for the

jth event, the distributed representation is initialized as $q_j = \frac{1}{l_j}\sum h^j_{l_j}$. Multi-head attention mechanism (MultiAttn) (Vaswani et al., 2017) is employed to softly select information from $H^{(M)}$ and get the representation of each event:

$$e_j = \text{MultiAttn}(q_j, H^{(M)}). \quad (3)$$

For brevity, we denote the vector representation of all events in X using a matrix E_X, where $E_X = \{e_1, e_2, e_3\} \in \mathbb{R}^{3\times d}$. Note that, through the embedding layer of RoBERTa, position information has been injected into the token representations. Therefore, E_X derived from token representations carries event order information. In addition, since E_X is obtained from the hidden states of RoBERTa, rich linguistic knowledge within RoBERTa could be utilized to enhance the comprehension of event semantics. By the same way, the representation of events within X' could be calculated, which we denote as $E_{X'}$.

4.2 Recognition Network

The recognition network models $q_\phi(z|X', A)$ based on $E_{X'}$ and A, where $E_{X'}$ is the representations of events within X'. Following traditional VAE, $q_\phi(z|X', A)$ is assumed to be a multivariate Gaussian distribution:

$$q_\phi(z|X', A) \sim N(\mu'(X', A), D), \quad (4)$$

where D denotes the identity matrix.

To obtain $\mu(X', A)$, we first combine $E_{X'}$ and adjacency matrix A using a GNN (Kipf et al., 2016):

$$E^{(U)'} = \sigma(AE_{X'}W^{(u)}). \quad (5)$$

where $\sigma(\cdot)$ is the sigmoid function; $W^{(u)} \in \mathbb{R}^{d\times d}$ is a weight matrix and $E^{(U)'}$ are relational information updated event representations.

Then a multi-head self-attention operation is performed to promote the fusion of event semantic information and relational information:

$$E^{(U)'} = \text{MultiAttn}(E^{(U)'}, E^{(U)'}). \quad (6)$$

Finally, to estimate $\mu(X', A)$, we aggregate information within $E^{(U)'}$ using a readout function $g(\cdot)$:

$$\mu' = g(E^{(U)'}). \quad (7)$$

Following Zhou et al. (2019) and Zhong et al. (2019), we set $g(\cdot)$ to be a mean-pooling operation.

Hence, by estimating μ' based on the *relational information updated* event representation $E^{(U)'}$, event graph knowledge about X' and A is involved into the latent variable z.

4.3 Prior Network

The prior network models $p_\theta(z|X)$ based on E_X, where E_X is the representation matrix of events in X. The same as the recognition network, $p_\theta(z|X)$ also follows multivariate normal distribution, while the parameters are different:

$$p_\theta(z|X) \sim N(\mu(X), D), \quad (8)$$

where D denotes the identity matrix.

To obtain $\mu(X)$, different from the recognition network, the prior network starts from updating E_X using a multi-head self-attention:

$$E^{(U)} = \text{MultiAttn}(E_X, E_X). \quad (9)$$

Then an additional multi-head self-attention operation is performed to get deeper representations:

$$E^{(U)} = \text{MultiAttn}(E^{(U)}, E^{(U)}). \quad (10)$$

Finally, $\mu(X)$ is estimated through aggregating information from $E^{(U)}$:

$$\mu = g(E^{(U)}), \quad (11)$$

where $g(\cdot)$ is a mean-pooling operation.

4.4 Merger

The merger module merges the latent variable z as well as updated (deep) representation of events into the Nth transformer layer of RoBERTa frame for predicting the relatedness score. To this end, we employ multi-head attention mechanism to softly select relevant information from z and $E^{(U)}$, and then update the hidden state of the Nth transformer layer of RoBERTa.

Specifically, in the pretraining stage:

$$H^{(N)^*} = \text{MultiAttn}(H^{(N)}, [\mu'; E^{(U)}]), \quad (12)$$

where $H^{(N)}$ is the hidden states of the Nth transformer layer of RoBERTa, and $H^{(N)^*}$ is the event graph information updated hidden states.

While in the finetuning and prediction stage:

$$H^{(N)^*} = \text{MultiAttn}(H^{(N)}, [\mu; E^{(U)}]). \quad (13)$$

Note that, given X, $p_\theta(\mu|X)$ achieves its maximum when $z = \mu$. Hence, making predictions based on μ could be regarded as finding the best explanation based on the most likely commonsense situation. Through integrating latent variable z, $H^{(N)^*}$ contains the event graph knowledge. By taking $H^{(N)^*}$ as the input of the subsequent $(N+1)$th transformer layers of RoBERTa for predicting the relatedness score, the abductive reasoning task is conducted based on the additional event graph knowledge.

4.5 Optimizing

The αNLI task requires model to choose a more likely hypothesis event from two candidates. However, in the pre-training stage, the negative examples are absent in the pseudo instances. To address this issue, following the method of Liu et al. (2019), in the pre-training stage ege-RoBERTa is trained to predict the masked tokens in the event sequence X rather than the relatedness score. In addition, in order to balance the masked token prediction loss with the KL term, we introduce an additional hyperparameter λ. Hence, the objective function in the pretraining stage is defined as follows:

$$\begin{aligned} L^{ELBO}(\theta,\phi) =& \mathbb{E}_{q(z|X',A)}\log L_{MLM}(X,z;\theta) \\ & - \lambda \mathrm{KL}(q_\phi(z|X',A)||p_\theta(z|X)), \end{aligned} \tag{14}$$

where $\log L_{MLM}(X,z;\theta)$ is the masked token prediction loss. Intuitively, through minimizing the KL term, we aim to transmit the event graph knowledge from the recognition network to the prior network.

In the finetuning stage, ege-RoBERTa is trained to adapt the learned event graph knowledge to the abductive reasoning task. Without the recogniton network, we formulate the objective function as:

$$L(\theta) = p_\theta(Y|z,X) = p_\theta(Y|z,X)p_\theta(z|X). \tag{15}$$

4.6 Training Details

We implement two different sizes of ege-RoBERTa model (i.e. ege-RoBERTa-base and ege-RoBERTa-large) based on RoBERTa-base framework and RoBERTa-large framework, respectively. For the ege-RoBERTa-base model, in the aggregator, the prior network, the recognition network and the merger, the dimension of the attention mechanism d is set as 768, and all multi-head attention layers contain 12 heads. While for the ege-RoBERTa-large model, d is equal to 1024 and all multi-head attention layers contain 16 heads. In the ege-RoBERTa-base model, *token* representations are aggregated from the 7th transformer layer of RoBERTa, and the latent variable is merged to the 10th transformer layer of RoBERTa. While for the ege-RoBERTa-large model, the aggregator and merger layer are set as the 14th and 20th layer, respectively. The balance coefficient λ equals 0.01. More details are provided in the Appendix.

5 Experiments

5.1 αNLI Dataset

The αNLI dataset (Bhagavatula et al., 2019) consists of 169,654, 1,532 and 4,056 $\langle O_1, O_2, H_1, H_2\rangle$ quadruples in training, development and test set, respectively. The observation events are collected from a short story corpus ROCstory (Mostafazadeh et al., 2016), while all of hypothesis events are independently generated through crowdsourcing.

(Posterior) Event Sequence	Story
Observed Event 1 ($\mathbf{O_1}$)	① I was doing exercise in gym.
Intermediary Event 1 ($\mathbf{I_1}$)	② I felt very hot.
Hypothesis Event ($\mathbf{H_1}$)	③ I got up to turn on the fan.
Intermediary Event 2 ($\mathbf{I_2}$)	④ The fan began to cool down my room.
Observed Event 2 ($\mathbf{O_2}$)	⑤ I felt much more comfortable.
A Pseudo Instance={X, X', A}, where $X = (O_1, H_1, O_2)$; $X' = (O_1, I_1, H_1, I_2, O_2)$ A is initialized from the event graph.	

Table 1: An example for illustrating the construction of pseudo instances used for pretraining ege-RoBERTa.

5.2 Construction of Event Graph

The event graph serves as an external knowledge base to provide information about the relationship between observation events and intermediary events. To this end, we build the event graph based on an auxiliary dataset, which are composed of two short story corpora independent to αNLI, i.e., VIST (Huang et al., 2016), and TimeTravel (Qin et al., 2019). Both VIST and TimeTravel are composed of five-sentences short stories. Totally there are 121,326 stories in the auxiliary dataset.

To construct the event graph, we define each sentence in the auxiliary dataset as a node in the event graph. To get the edge weight W_{ij} between two nodes V_i and V_j (i.e., the probability that V_j is the subsequent event of V_i), we finetune a RoBERTa-large model through a next sentence prediction task. Specifically, we define adjacent sentence pairs in the story text (for example, [$1st$, $2nd$] sentence, [$4th$, $5th$] sentence of a story) as positive instances, define nonadjacent sentence pairs or sentences pairs in reverse order (such as [$1st$, $3rd$] sentence, [$5th$, $4th$] sentence of a story) as negative instances. After that we sample 300,000 positive and 300,000 negative instances from the auxiliary dataset. Then given an event pair (V_i, V_j), the finetuned RoBERTa-large model would be able to predict the probability that V_j is the subsequent event of V_i.

Event Graph Based Pseudo Instance Set for Pretraining ege-RoBERTa To effectively utilize the event graph knowledge, we induce a set of pseudo instances for pretraining the ege-RoBERTa model. Specifically, given a five-sentence-story within the auxiliary dataset, as Table 1 shows, we define the 1st and 5th sentence of the story as two

observed events, the 3rd sentence as the hypothesis event, the 2nd and 4th sentence as intermediary events, respectively. In this way, the posterior event sequence X' and the event sequence X of a pseudo instance could be obtained. In addition, given X', we initialize the elements of the adjacency matrix A using the edge weights of the event graph, and scale A so that its row sums equal to 1. After the above operations, each pseudo instance is composed of an event sequence X, a posterior event sequence X' which contains intermediary event information, and an adjacency matrix A which describes relationships between events within X'.

5.3 Baselines

We compare ege-RoBERTa with:

• SVM uses features about length, overlap and sentiment to predict the more likely hypothesis event.

• Infersent (Conneau et al., 2017) represents sentences using a Bi-LSTM, and predicts the relatedness score using MLP.

• GPT (Radford et al., 2018) is a multilayer-transformer based unidirectional pretrained language model.

• BERT (Devlin et al., 2019) is a multilayer-transformer based bi-directional pretrained language model.

• RoBERTa (Liu et al., 2019) refers robustly optimized BERT.

• ege-RoBERTa$_{\text{u(npretrained)}}$ refers to the ege-RoBERTa model without the pretraining stage.

• ege-RoBERTa$_{\lambda=0}$ refers to setting the balance coefficient to 0 in the pretraining stage. Note that all pretrained-language-model-based baselines (i.e., GPT, BERT and RoBERTa) are finetuned on the αNLI dataset as the method of Bhagavatula et al. (2019) to adapt to the abductive reasoning task.

In addition, we also list two concurrent works: (i) L^{2}R (Zhu et al., 2020) learns to rank the candidate hypotheses with a novel scoring function. (ii) RoBERTa-GPT-MHKA (Paul et al., 2020) enhances pretrained language model with social and causal commonsense knowledge for αNLI task.

5.4 Quantitative Analysis

We list the prediction accuracy (%) in Table 2, and observe that:

(1) Compared with SVM and Infersent, pretrained language model based methods: GPT, BERT, RoBERTa and ege-RoBERTa show significant better performances in abductive reasoning task. This is because through the pre-training stage language models could capture rich linguistic knowledge that is helpful for understanding the semantics of events.

Methods	**Accu. (%)**
SVM	50.6
Infersent (Conneau et al., 2017)	50.8
GPT (Radford et al., 2018)	63.1
BERT-base (Devlin et al., 2019)	63.3
RoBERTa-base (Liu et al., 2019)	71.5
BERT-large (Devlin et al., 2019)	68.9
RoBERTa-large (Liu et al., 2019)	83.9
Concurrent Methods	
L^{2}R (Zhu et al., 2020)	86.8
RoBERTa-GPT-MHKA (Paul et al., 2020)	87.1
This Work	
ege-RoBERTa-large$_{\text{u}}$	83.8
ege-RoBERTa-large$_{\lambda=0}$	84.2
ege-RoBERTa-base	**75.9**
ege-RoBERTa-large	**87.5**
Human Performance	91.4

Table 2: Accuracy on the test set of αNLI.

(2) Comparison between ege-RoBERTa-large$_{\text{u}}$ with ege-RoBERTa-large shows that the pre-training process can increase the accuracy of abductive reasoning. In addition, comparison between ege-RoBERTa-large$_{\lambda=0}$ with ege-RoBERTa-large indicates that in the pre-training process, ege-RoBERTa could capture the event graph knowledge through the latent variable to enhance the abductive reasoning. Furthermore, the relative close performance between ege-RoBERTa-large$_{\text{u}}$ and ege-RoBERTa-large$_{\lambda=0}$ suggest that the main improvements of the performance is brought by the event graph knowledge.

(3) Compared to RoBERTa, ege-RoBERTa achieves higher prediction accuracy for both the base and large sized model. This result confirms our motivation that learning event graph knowledge could be helpful for the abductive reasoning task.

(4) According to Bhagavatula et al. (2019), human performance on the test set of αNLI is 91.4%. While the RoBERTa-large model has achieved an accuracy of 83.9%. Therefore, further improvements over RoBERTa-large could be challenging. Through learning the event graph knowledge, our proposed method ege-RoBERTa further improves the relative accuracy.

(5) Our approach has comparable performance with the SOTA concurrent work, which combines RoBERTa with GPT, and incorporates social and causal commonsense into model. The combination of both methods would further increase the model performance.

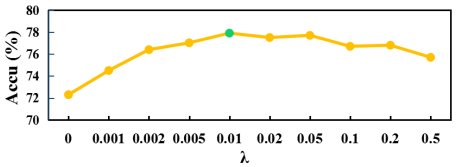

Figure 4: Accuracy of ege-RoBERTa-base pretrained with different balance coefficient λ.

Model	Accuracy (%)
ege-RoBERTa-base	**77.9**
-w/ $\tilde{A}$	75.5
-w/ $\tilde{I}_1$ and $\tilde{I}_2$	76.0

Table 3: Prediction accuracy of the ege-RoBERTa-base model pretrained with randomly initialized adjacency matrix $\tilde{A}$ / randomly sampled intermediary events $\{\tilde{I}_1, \tilde{I}_2\}$.

5.5 Ablation Study

All studies are conducted on the development set of the αNLI using the ege-RoBERTa-base model.

Influence of the Balance Coefficient In the pretraining stage, the balance coefficient λ controls the trade off between event graph knowledge learning and abductive reasoning. To investigate the specific influence of the balance coefficient, we compare the performance of ege-RoBERTa model pretrained with different λ. As shown in Figure 4, the prediction accuracy continues to increase as λ increases from 0 to 0.01. This is because adequate event graph knowledge can offer guidance for the abductive reasoning task. While when λ exceeds 0.05, the accuracy start to decrease, as the over-emphasis of event graph knowledge learning would in turn undermine the model performance.

Influence of the External Commonsense Knowledge We study the specific effect of the event relational information and the intermediary event information by controlling the generation of pseudo instances. In specific, we eliminate the influence of the adjacency matrix A by replacing A with a randomly initialized matrix $\tilde{A}$. Similarly, the influence of the intermediary events I_1 and I_2 is eliminated through substituting them by two randomly sampled events $\tilde{I}_1$ and $\tilde{I}_2$. As Table 3 shows, both the replacement of A and $\{I_1, I_2\}$ lead to obvious decrease of model performance. This demonstrates that ege-RoBERTa can use both two kinds of event graph knowledge for enhancing the abductive reasoning task.

5.6 Sensitivity Analysis

To find out if the improvement of Ege-RoBERTa is brought by a certain dataset, and the specific relationship between the model performance with the number of pseudo instances, we conduct following experiments: (1) excluding a certain dataset when inducing pseudo instances; (2) pretraining the ege-RoBERTa-base model with different number of pseudo instances. The corresponding results on the dev set of αNLI is shown in Table 4.

Dataset	-w/o TimeTravel		-w/o VIST	
Accuracy	76.6		75.7	
#Pseudo Instances	40,000	60,000	80,000	100,000
Accuracy	74.3	75.4	76.2	77.0

Table 4: Sensitivity analysis about the source and number of pseudo instances on the dev set of αNLI.

Model	**Posterior event Sequence**	**Accu.**
RoBERTa	—	73.2
ege-RoBERTa	$X' = \{O_1, \boldsymbol{I_1}, H_i, O_2\}$	77.1
	$X' = \{O_1, H_i, \boldsymbol{I_1}, O_2\}$	76.3
	$X' = \{O_1, \boldsymbol{I_1}, \boldsymbol{I_2}, H_i, O_2\}$	76.6
	$X' = \{O_1, H_i, \boldsymbol{I_1}, \boldsymbol{I_2}, O_2\}$	75.8
	$X' = \{O_1, \boldsymbol{I_1}, H_i, \boldsymbol{I_2}, O_2\}$	**77.9**

Table 6: Prediction accuracy (%) of the ege-RoBERTa-base model pretrained with different forms of posterior event sequence.

We can find that, the elimination of both dataset leads to decrease of model performances. This suggests that the ege-RoBERTa model could capture relevant event graph knowledge from both dataset. While the prediction accuracy continues to increase along with the number of pseudo instances used for pretraining the ege-RoBERTa model. This is because the accumulation of commonsense knowledge is helpful for the abductive reasoning task. In addition, it also indicates that the model performance could be further improved if the auxiliary dataset is even more enlarged.

5.7 Case study

Table 5 provides an example of model prediction results. Given two observed events O_1 "*hates Fall*" and O_2 "*didn't have to experience Fall in Guam*", the hypothesis event H_1 "*moved to Guam*" is more likely to explain the two motivations of observed events. However, H_1 implicitly relies on a precondition that in Guam, Fall could be eluded. Correspondingly, in the auxiliary dataset, there is information supporting the hypothesis event H_1 that there is no Fall in Guam. In this case, ege-RoBERTa chooses the hypothesis event H_1, whereas RoBERTa chooses the wrong hypothesis event H_2. This indicates that ege-RoBERTa could learn the event graph knowledge in the pretraining process for improving the reasoning performance.

Observed Events	Hypothesis Events	Model	Commonsense Knowledge from EG.
O_1: I hated Fall. O_2: I became happier because I didn't have to experience Fall in Guam.	H_1: I moved to Guam. ($\surd$)	ege-RoBERTa	I_1: It reminded me of death. H: I couldn't stand Fall so I decided to move.
	H_2:I took a vacation during the Fall. ($\times$)	RoBERTa	I_2: I moved to Guam where there was no Fall season.

Table 5: Example of abductive reasoning result made by RoBERTa and ege-RoBERTa, respectively.

6 Discussion

In this paper, to involve the event graph knowledge, we formalize the posterior event sequence as $X' = \{O_1, I_1, H_i, I_2, O_2\}$. While our approach also allows other forms of posterior event sequences, such as $X' = \{O_1, H_i, I_1, O_2\}$, $X' = \{O_1, I_1, H_i, O_2\}$, or $X' = \{O_1, I_1, I_2, H_i, O_2\}$, etc. We also pretrained ege-RoBERTa on pseudo-instance sets derived by these manners. The results are shown in Table 6. We find that whatever forms of posterior event sequences involved in ege-RoBERTa, our approach can achieve consistently better performance than the baseline method. This confirms that our approach is sufficiently generalizable to deal with various forms of external event-sequence knowledge. Furthermore, ege-RoBERTa can also be equipped with more types of event graph knowledge, such as background knowledge by: formalizing the posterior event sequence as $X' = \{B_1, \ldots, B_m, E_1, \ldots, E_n\}$, where $\{B_1, \ldots, B_m\}$ is a set of background events for a given prior event sequence $\{E_1, \ldots, E_n\}$. This demonstrates the potential of ege-RoBERTa in learning different kinds of event graph knowledge for different event inference tasks.

7 Related Work

7.1 Abductive Reasoning

Most previous studies focus on formal logic based abductive reasoning (Eshghi et al., 1988; Levesque, 1989; Konolige, 1990; Paul, 1993). To infer the most reasonable hypothesis, the abductive reasoning process could be divided into two steps: (1) proposing reasonable hypotheses; (2) finding the best explanation from the hypotheses (Levesque, 1989; Konolige, 1990; Paul, 1993).

However, the rigidity of formal logic limits its application in NLP domain. To facilitate this, Bhagavatula et al. (2019) proposed a text based abductive reasoning task αNLI. To solve the this task, Zhu et al. (2020) formalize αNLI as a rank learning task, and propose a novel ranking function. While Paul et al. (2020) enhances the reasoning model with social commonsense and causal commonsense knowledge. Compared to their works, for enhancing the abductive reasoning process, we propose to incorporate event graph knowledge by a CVAE based model ege-RoBERTa. In addition, we argue that our approach can be easily extended to other event inference tasks.

7.2 Event Graph Based Natural Language Inference

Understanding events and their relationships are crucial for various natural language inference (NLI) tasks (Kruengkrai et al., 2017). Hence, a number of previous studies explore conducting NLI tasks based on event graphs.

For example, to predict the subsequent event for a given event context, Li et al. (2018) build an event evolutionary graph (EEG), and make prediction using a scaled graph neural network. While Wu et al. (2019) predict the propagation of news event through combining an historical event propagation graph with temporal point process. In addition to the event prediction related tasks, Liu et al. (2017) propose to enhance the news recommendation by incorporating additional event graph information. Liu et al. (2016) detect the textual contradiction by using event graphs as additional evidence.

In this paper, we employ event graph knowledge for guiding the abductive reasoning. To this end, we propose a variational autoencoder based framework ege-RoBERTa, which employs a latent variable z to implicitly capture the necessary event graph knowledge and enhance the pretrained language model RoBERTa.

8 Conclusion

In this paper, we propose a variational autoencoder based framework ege-RoBERTa with a two-stage training procedure for the abductive reasoning task. In the pretraining stage, ege-RoBERTa is able to learn commonsense knowledge from an event graph through the latent variable, then in the following stage the learned event graph knowledge can be adapted to the abductive reasoning task. Experimental results show improvement over the baselines on the αNLI task.

9 Acknowledgments

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Key Research and Development Program of China (2020AAA0106501), and the National Natural Science Foundation of China (61976073).

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Eugene Charniak and Solomon Eyal Shimony. 1990. *Probabilistic semantics for cost based abduction*. Brown University, Department of Computer Science.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 NACCL*, pages 4171–4186.

K Eshghi, , and RA Kowalski. 1988. Abduction through deduction. *Logic Programming Section Technical Report, Department of Computing, Imperial College, London*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Thomas N. Kipf, Thomas N. Welling, Max, Thomas N. Welling, Max, and Max Welling. 2016. Semi-supervised classification with graph convolutional networks.

Kurt Konolige. 1990. Closure+ minimization implies abduction. In *Proceedings of PRICAI90*.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Hector J Levesque. 1989. A knowledge-level account of abduction. In *Proceedings of the 11th international joint conference on Artificial intelligence-Volume 2*, pages 1061–1067.

Zhongyang Li, Xiao Ding, , and Ting , Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4201–4207.

Maofu Liu, Limin Wang, Liqiang Nie, Jianhua Dai, and Donghong Ji. 2016. Event graph based contradiction recognition from big data collection. *Neurocomputing*, 181:64–75.

Shenghao Liu, Bang Wang, and Minghua Xu. 2017. Event recommendation based on graph random walking and history preference reranking. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 861–864.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*, pages 839–849.

Hwee Tou Ng, , , and Raymond J Mooney. 1990. The role of coherence in constructing and evaluating abductive explanations. In *Working Notes, AAAI Spring Symposium on Automated Abduction, Stanford, California*.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra-and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743.

Debjit Paul, Anette Frank, and and . 2020. Social commonsense reasoning with multi-head knowledge attention. *arXiv preprint arXiv:2010.05587*.

Gabriele Paul. 1993. Approaches to abductive reasoning: an overview. *Artificial intelligence review*, 7(2):109–152.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL*

https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.

Kihyuk Sohn, Honglak Lee, Xinchen Yan, et al. 2015. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Weichang Wu, Huanxi Liu, Xiaohu Zhang, Yu Liu, and Hongyuan Zha. 2019. Modeling event propagation via graph biased temporal point process. *arXiv preprint arXiv:1908.01623*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745*.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*.

Yunchang Zhu, Liang Pang, Yanyan Lan, and Xueqi Cheng. 2020. L2r^2: Leveraging ranking for abductive reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1961–1964.