

Modeling Event Background for If-Then Commonsense Reasoning Using Context-aware Variational Autoencoder

Li Du, Xiao Ding, Ting Liu* and Zhongyang Li

Research Center for Social Computing and Information Retrieval

Harbin Institute of Technology, China

{ldu, xding, tliu, zyli}@ir.hit.edu.cn

Abstract

Understanding event and event-centered commonsense reasoning are crucial for natural language processing (NLP). Given an observed event, it is trivial for human to infer its intents and effects, while this type of If-Then reasoning still remains challenging for NLP systems. To facilitate this, a If-Then commonsense reasoning dataset Atomic is proposed, together with an RNN-based Seq2Seq model to conduct such reasoning. However, two fundamental problems still need to be addressed: first, the intents of an event may be multiple, while the generations of RNN-based Seq2Seq models are always semantically close; second, external knowledge of the event background may be necessary for understanding events and conducting the If-Then reasoning. To address these issues, we propose a novel context-aware variational autoencoder effectively learning event background information to guide the If-Then reasoning. Experimental results show that our approach improves the accuracy and diversity of inferences compared with state-of-the-art baseline methods.

1 Introduction

Recently, event-centered commonsense knowledge has attracted much attention (Chambers and Jurafsky, 2008; Segers et al., 2016; Wang et al., 2017; Li et al., 2018), because of understanding events is an important component of NLP. Given a daily-life event, human can easily understand it and reason about its causes, effects, and so on. However, it still remains a challenging task for NLP systems. This is partly due to most of them are trained for task-specific datasets or objectives, which results in models that are adapt at finding task-specific underlying correlation patterns but have limited capability in simple and explainable commonsense reasoning (Sap et al., 2018).

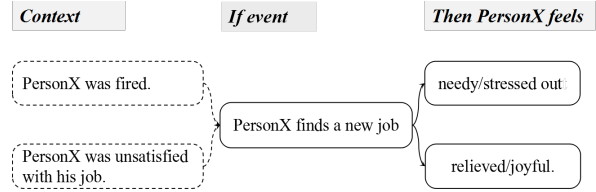


Figure 1: A illustration of two challenging problems in If-Then reasoning. (a) Given an observed event, the feelings about this event could be multiple. (b) Background knowledge is need for generating reasonable inferences, which is absent in the dataset (marked by dashed lines).

To facilitate this, Rashkin et al. (2018) build the Event2Mind dataset and Sap et al. (2018) present the Atomic dataset, mainly focus on nine If-Then reasoning types to describe causes, effects, intents and participant characteristic about events. Together with these datasets, a simple RNN-based encoder-decoder framework is proposed to conduct the If-Then reasoning.

However, there still remains two challenging problems. First, as illustrated in Figure 1, given an event “PersonX finds a new job”, the plausible feeling of PersonX about that event could be multiple (such as “needy/stressed out” and “relieved/joyful”). Previous work showed that for the one-to-many problem, conventional RNN-based encoder-decoder models tend to generate generic responses, rather than meaningful and specific answers (Li et al., 2016; Serban et al., 2016).

Second, as a commonsense reasoning problem, rich background knowledge is necessary for generating reasonable inferences. For example, as shown in Figure 1, the feeling of PersonX upon the event “PersonX finds a new job” could be multiple. However, after given a context “PersonX was fired”, the plausible inferences would be narrowed down to “needy” or “stressed out”.

To better solve these problems, we propose a context-aware variational autoencoder (CWVAE) together with a two-stage training procedure. Vari-

*Corresponding author

ational Autoencoder (VAE) based models have shown great potential in modeling the one-to-many problem and generate diversified inferences (Bowman et al., 2015; Zhao et al., 2017).

In addition to the traditional VAE structure, we introduces an extra context-aware latent variable in CWVAE to learn the event background knowledge. In the pretrain stage, CWVAE is trained on an auxiliary dataset (consists of three narrative story corpora and contains rich event background knowledge), to learn the event background information by using the context-aware latent variable. Subsequently, in the finetune stage, CWVAE is trained on the task-specific dataset to adapt the event background information to each specific aspect of If-Then inferential target (e.g., intents, reactions, etc.).

Experiments on the Event2Mind and Atomic dataset show that our proposed approach outperforms baseline methods in both the accuracy and diversity of inferences. The code is released at <https://github.com/sjcfr/CWVAE>.

2 Background

Before specifically describing two dataset — Event2Mind and Atomic used in this paper as well as the If-Then reasoning task, for clarity, we define the following terminologies:

Base event: the prerequisite event in If-Then reasoning, organized as a verb phrase with a predicate and its arguments, such as the event “PersonX finds a new job” shown in Figure 1.

Inference dimension: a particular If-Then reasoning type, e.g., intents, effects of the base event. Details are shown in Table 1 and Table 2.

Target: the inferential results. For example, as shown in Figure 1, given a base event “PersonX finds a new job” and one inference dimension “xReact”, the targets could be “*relieved*” or “*needy*”. Notice that each inference dimension can have multiple targets.

Event2Mind Dataset contains 25K base events and 300K targets, annotated through crowdsourcing. Event2Mind is organized in a hierarchical form: each base event has three types of inference dimensions, and given a base event, under one of inference dimensions, several targets may simultaneously exist. Table 1 shows the (base event-inference dimension-target) hierarchical structure through an example from Event2Mind.

Atomic Dataset Inspired by Event2Mind, the

Base event	Inference Dim.	Target
PersonX writes PersonY a letter	xIntent	to send a message, express themselves
	xReact	nervous, thoughtful
	oReact	indifferent, receptive

Table 1: Hierarchical structure of Event2Mind dataset. For specific inference dimensions, “x” and “o” refers to PersonX and others respectively.

Base event	Inference Dim.	Target
PersonX adopts a child	xIntent	to help another person, to have a child
	xNeed	to visit adoption agency, to be approved for adoption
	xAttr	compassionate, generous
	xEffect	becomes a parent, gains love and companionship
	xWant	take child home, buy child clothes
	xReact	happy, caring
	oReact	has a parent, receives love and affection
	oWant	try on new clothes, to have a family
	oEffect	has a parent, Receives love and affection

Table 2: Hierarchical structure of Atomic dataset. For specific inference dimensions, “x” and “o” refers to PersonX and others respectively.

Atomic dataset shares the same hierarchical structure as Event2Mind, while scales up the size of dataset and expands the scope to nine types of inference dimensions. Table 2 shows the (base event-inference dimension-target) hierarchical structure through an example from Atomic. Though Atomic covers the inference dimensions of Event2Mind, the base event collection of Event2Mind is nonidentical to that of Atomic.

Problem Definition The If-Then reasoning task could be formally defined as a conditional one-to-many generation problem: given a base event x and one inference dimension d , the model is required to generate targets $y = f(x, d)$ as close to the ground truths as possible. Both x and y consist of sequence of words: $x = \{x_1, \dots, x_m\}$, and $y = \{y_1, \dots, y_n\}$, where m and n denotes the length of x and y , respectively.

Conditional Variational Autoencoder The variational autoencoder (VAE) defines a generative framework suited for one-to-many generation problem (Kingma and Welling, 2014). While con-

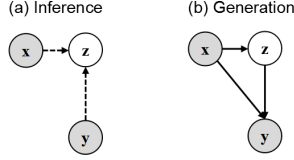


Figure 2: Illustration of inference and generation process of CVAE in a directed graph. Dashed lines represent the inference of z . Solid lines represent the generation process.

ditional variational autoencoder (CVAE) (Sohn et al., 2015) is an extension of VAE on the conditional generation problem. As shown in Figure 2 (a), CVAE characterizes the conditional one-to-many generation problem using three random variables: event x , target y and a latent variable z , which is used for modeling the latent distribution of semantic over targets given an event. Hence, under a certain inference dimension, with regard to the latent semantic variable z , the conditional generation problem could be expressed as $p(y|x) = \int p(y|x, z)p(z|x)dz$. CVAE models $p(y|x, z)$ and $p(z|x)$ using deep neural networks (parameterized by θ) $p_\theta(y|x, z)$ and $p_\theta(z|x)$. Then as illustrated in Figure 2 (b), y could be generated from x and z .

CVAE is trained to maximize the conditional likelihood $p(y|x)$, which involves an intractable marginalization over the latent variable z . Instead, following Kingma and Welling (2014), a practical way is to introduce another deep network (parameterized by ϕ) $q_\phi(z|x, y)$ to approximate the true posterior distribution $p(z|x, y)$ and maximize the evidence lower bound (ELBO) of the log-likelihood function:

$$L^{ELBO}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x, y)} \log(p_\theta(y|x, z)) - \text{KL}(q_\phi(z|x, y) || p_\theta(z|x)) \leq \log p(y|x) \quad (1)$$

Therefore, CVAE is composed of three neural networks in general. We refer to $p_\theta(z|x)$ as a prior network, $q_\phi(z|x, y)$ as a recognition network, and $p_\theta(y|x, z)$ as a neural decoder.

3 Context-aware Variational Autoencoder

Traditional CVAE can model the event-target relation. In other words, given an observed event, CVAE can generate its corresponding targets. While in this paper we model the If-Then reasoning as a $[(\text{background}), \text{event}]$ -target process. It means that in addition to the observed

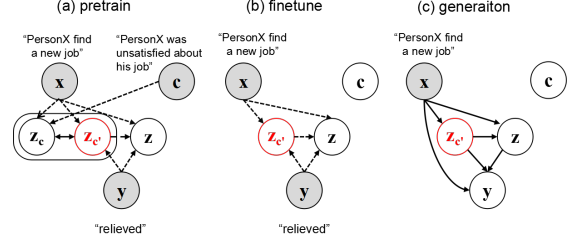


Figure 3: Illustration of pretrain, finetune and generation process of CWVAE in a directed graph. Dashed lines represent the inference of z , z_c and $z_{c'}$. Solid lines represent the generation process. Red circle denotes the context-aware latent variable.

event, we also want to involve the event background knowledge (which can be learned from event contexts) to generate the reasonable targets.

To this end, we propose a context-aware variational autoencoder (CWVAE), with two additional latent variables: a context-acquiring latent variable z_c to directly acquire context information, and a context-aware latent variable $z_{c'}$ to learn background knowledge from z_c , as shown in Figure 3 (a). However, the event context information is absent in the Event2Mind and Atomic dataset. To learn from the external event context information, we design the following two-stage training procedure for CWVAE.

Pretrain: Learning Event Background Knowledge from Auxiliary Dataset In the pretrain stage, CWVAE is trained on three narrative story corpora with rich event context information. As shown in Figure 3 (a), context-acquiring latent variable z_c is directly conditioned on the context c . Hence, z_c could be employed for acquiring background knowledge from event contexts. Then, we minimize the distance between z_c and the context-aware latent variable $z_{c'}$, by which the event background knowledge is transferred from z_c to $z_{c'}$.

Finetune: Adapt Event Background Knowledge to Each Inference Dimension In the finetune stage, as shown in Figure 3 (b), CWVAE is trained on the Event2Mind and Atomic dataset without the event context information. Pretrained CWVAE is finetuned to learn the specific inferential knowledge of each inference dimension. After the training procedure, as shown in Figure 3 (c), samples of z is generated based on x and samples of $z_{c'}$, where $z_{c'}$ contains rich event background knowledge helpful for If-Then reasoning.

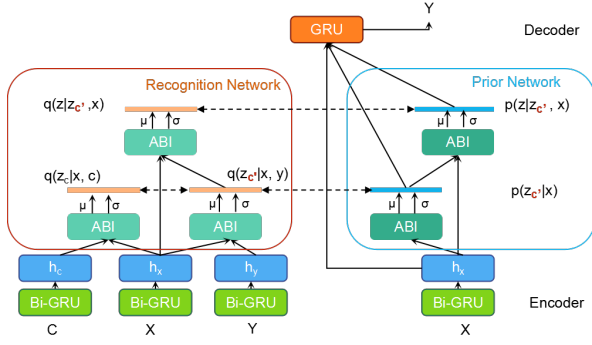


Figure 4: Architecture of CWVAE. We mark Neural encoder in green, prior network in blue, recognition network in brown and neural decoder in orange, respectively.

3.1 Architecture of CWVAE

As shown in Figure 4, CWVAE is mainly composed of four parts: a neural encoder that provides distributed representations of base events/targets, a recognition network for inferring $q_\phi(z_{c'}|x, y)$, $q_\phi(z_c|x, c)$ and $q_\phi(z|z_{c'}, x)$, a prior network for modeling $p_\theta(z_{c'}|x)$ and $p_\theta(z|x, z_{c'})$, and a neural decoder that integrates the information from z and $z_{c'}$ to generate targets.

Neural Encoder We employ a bidirectional GRU as neural encoder, which encodes context c , event x and target y into distributed representations $h^c = \{h_1^c, \dots, h_{l_c}^c\}$, $h^x = \{h_1^x, \dots, h_{l_x}^x\}$ and $h^y = \{h_1^y, \dots, h_{l_y}^y\}$, where l_c , l_x and l_y is the length of c , x and y , respectively.

Recognition Network The recognition network models $q_\phi(z|x, y)$, $q_\phi(z_c|x, c)$, $q_\phi(z|z_{c'}, x)$ based on h^x , h^y and h^c .

Following traditional VAE, the above-mentioned three distributions are assumed to be multivariate Gaussian distribution with a diagonal covariance structure:

$$\begin{aligned} q_\phi(z_c|x, c) &\sim N(\mu_{z_c}(x, c), \sigma_{z_c}(x, c)I) \\ q_\phi(z_{c'}|x, y) &\sim N(\mu_{z_{c'}}(x, y), \sigma_{z_{c'}}(x, y)I) \\ q_\phi(z|x, y) &\sim N(\mu_z(x, y), \sigma_z(x, y)I) \end{aligned} \quad (2)$$

where μ denotes the mean of the distribution, σ denotes the standard deviation of the distribution, and I denotes the identity matrix.

Given h^x , h^y and h^c , we propose a novel attention-based inferer (ABI) module to estimate the mean and standard deviation of $q_\phi(z_c|x, c)$, $q_\phi(z_{c'}|x, y)$ and $q_\phi(z|x, y)$:

$$\begin{aligned} \mu_{z_c}, \sigma_{z_c} &= \text{ABI}_c(h^c, h^x) \\ \mu_{z_{c'}}, \sigma_{z_{c'}} &= \text{ABI}_{c'}(h^y, h^x) \\ \mu_z, \sigma_z &= \text{ABI}_z(z_{c'}, h^x) \end{aligned} \quad (3)$$

Briefly, through the attention mechanism, ABI can capture the semantic interaction between input sequences, and estimate the parameters of distributions based on it. We will introduce the specific structure of ABI in below.

Prior Network Prior Network models $p_\theta(z_{c'}|x)$ and $p_\theta(z|x, z_{c'})$ based on h^x . The distribution of $p_\theta(z_{c'}|x)$ and $p_\theta(z|x, z_{c'})$ are still assumed to be multivariate Gaussian, whereas the parameters are different:

$$\begin{aligned} p_\theta(z|x) &\sim N(\mu'_z(x), \sigma'_z(x)I) \\ p_\theta(z_{c'}|x) &\sim N(\mu'_{z_{c'}}(x), \sigma'_{z_{c'}}(x)I) \end{aligned} \quad (4)$$

where μ' denotes the mean of the distribution, σ' denotes the standard deviation of the distribution and I denotes the identity matrix.

Then the attention-based inferer module is still employed to estimate parameters of distributions:

$$\begin{aligned} \mu'_{z_{c'}}, \sigma'_{z_{c'}} &= \text{ABI}'_{c'}(h^x, h^x) \\ \mu'_z, \sigma'_z &= \text{ABI}'_z(z_{c'}, h^x) \end{aligned} \quad (5)$$

Neural Decoder Given the base event x , the semantic latent variable z , and the context-aware latent variable $z_{c'}$, the neural decoder defines the generation probability of y as following:

$$p(y|x, z, z_{c'}) = \prod_{j=1}^n p(y_j|y < j, z, z_{c'}, x) \quad (6)$$

where $p(y_j|y < j, z, z_{c'}, x) = g(y_{j-1}, s_{j-1}, e_j)$, $g(\cdot)$ is an attention-based feed forward model, $e_j = \sum_i \alpha_{ji} h_i^x$ is the context vector and s_{j-1} is the hidden state of the decoder. We obtain $g(\cdot)$ and e_j the same way as Bahdanau et al. (2014). Whereas our decoder differs from Bahdanau et al. (2014) in that our model integrates the context-aware latent variable $z_{c'}$ and semantic latent variable z in the computation of $s_j = \text{GRU}([E_{y_j}; s_{j-1}, z, z_{c'}])$, where E_{y_j} is the word embeddings of target words.

Note that through concatenating z and $z_{c'}$ with E_{y_j} and s_{j-1} , s_j could be affected by context-aware latent variable $z_{c'}$ and semantic latent variable z . This allows model to directly access to the event background knowledge from $z_{c'}$. In addition, the randomness of z and $z_{c'}$ would increase the diversity of model generation.

Attention-based Inferer Attention mechanism has shown strong ability in capturing semantic interactions (Gong et al., 2017). Inspired by the co-attention mechanism (Parikh et al., 2016), we propose an attention-based inferer (ABI) to estimate

the mean and standard deviation of a distribution belongs to $p_\theta(\cdot)$ or $q_\phi(\cdot)$ by capturing semantic interactions of input sequences.

Specifically, given two input sequences (e.g., representations of contexts and events) $a = \{a_1, \dots, a_{l_a}\}$ and $b = \{b_1, \dots, b_{l_b}\}$ with length l_a and l_b , we first obtain the attention scores from each side through:

$$\begin{aligned}\gamma_t^a &= \frac{\exp[(W_a h_t^a)^T (W_b h_t^b)]}{\sum_i \exp[(W_a h_i^a)^T (W_b h_i^b)]} \\ \gamma_t^b &= \frac{\exp[(W_a h_t^a)^T (W_b h_t^b)]}{\sum_i \exp[(W_a h_i^a)^T (W_b h_i^b)]}\end{aligned}\quad (7)$$

where $W_a \in \mathbb{R}^{d \times d_a}$ and $W_b \in \mathbb{R}^{d \times d_b}$ are parameter weights.

With these attention scores, the context vectors of both sequences are given by:

$$\begin{aligned}c_t^a &= a \gamma_t^a \\ c_t^b &= b \gamma_t^b\end{aligned}\quad (8)$$

Then we perform a mean pooling operation on context vectors of both sequences:

$$\begin{aligned}\bar{c}^a &= \frac{1}{l_a} \sum_{t=1}^{l_a} c_t^a \\ \bar{c}^b &= \frac{1}{l_b} \sum_{t=1}^{l_b} c_t^b\end{aligned}\quad (9)$$

To obtain the mean and standard deviation, the pooled context vectors \bar{c}^a and \bar{c}^b which carry semantic interaction between two sequences, are concatenated and projected into a latent semantic space through a nonlinear transformation:

$$h_z = \tanh(W[\bar{c}^a; \bar{c}^b] + b_z) \quad (10)$$

Finally the mean and standard deviation are generated through a nonlinear transformation over h_z :

$$\begin{aligned}\mu &= W_\mu h_z + b_\mu \\ \sigma &= \text{softplus}(W_\sigma h_z + b_\sigma)\end{aligned}\quad (11)$$

3.2 Optimizing

With the incorporation of $z_{c'}$, the original loglikelihood could be decomposed as:

$$\log p(y|x) = \iint p(y|z, z_{c'}, x) dz dz_{c'} \quad (12)$$

Then following traditional CVAE, the ELBO of CWVAE is defined as follows:

$$\begin{aligned}L^{ELBO}(\theta, \phi) &= \underbrace{\mathbb{E}_{q_\phi(z|x, z_{c'}) q_\phi(z_{c'}|x, y)} \log p_\theta(y|x, z, z_{c'})}_{\text{Reconstruction Loss}} \\ &\quad - \underbrace{\int q_\phi(z|x, y) \text{KL}(q_\phi(z_{c'}|x, y) || p_\theta(z_{c'}|x)) dz}_{\text{KL term}} \\ &\quad - \underbrace{\int q_\phi(z_{c'}|x, y) \text{KL}(q_\phi(z|x, z_{c'}) || p_\theta(z|x, z_{c'})) dz_{c'}}_{\text{KL term}}\end{aligned}\quad (13)$$

which is the objective function at the finetune stage.

While in the pretrain stage, as we aim to learn background knowledge through minimizing the distance between z_c and $z_{c'}$, in addition to L^{ELBO} , a context-aware regulation term is introduced:

$$\begin{aligned}L^{PT}(\theta, \phi) &= L^{ELBO}(\theta, \phi) \\ &\quad - \lambda \underbrace{\int \text{KL}(q_\phi(z_c|x, c) || q_\phi(z_{c'}|x, y)) dz_{c'}}_{\text{context-aware regulation}}\end{aligned}\quad (14)$$

where the context aware regularization term is the KL distance between z and $z_{c'}$. Through minimizing the context aware regularization term, we aim to pass event context knowledge from z_c to the context aware latent variable $z_{c'}$.

3.3 Training Details

To test the performance of CWVAE, we split the Event2Mind and Atomic dataset into training, development and test sets (80%, 10%, 10%) in the same way as [Rashkin et al. \(2018\)](#) and [Sap et al. \(2018\)](#), respectively.

We initialize the embedding layer from 300d GloVe word embeddings. The neural encoder is chosen to be biGRU with 300 hidden units. For the ABI module, size of W_a and W_b is set to be $100 \times d_a$ and $100 \times d_b$ respectively. The dimension of z_c , $z_{c'}$ and z is all set as 40. The neural decoder is set to be GRU with 300d hidden state. Regulation coefficient λ of context-aware regulation term is set to be 0.1. Models are trained using an Adam optimizer ([Kinga and Adam, 2015](#)) with a learning rate of 0.001.

4 Experiments

4.1 Auxiliary Dataset

The auxiliary dataset is built upon three human-written story corpora: ROCStories ([Mostafazadeh](#)

Context	Event	Inference Target
① jason had been really stressed out at work. ② he decided he needed a different kind of job. ③ jason applied for a job in a different field.	④ he got the job .	⑤ jason was much happier at his new job .

Table 3: An example for the construction of auxiliary dataset. For a five-sentence-paragraph, the first three sentences are taken as event context, while the fourth and fifth sentence is taken as base event and target respectively.

et al., 2016), VIST (Huang et al., 2016) and WritingPrompts (Fan et al., 2018). ROCStories and VIST are composed of short stories with five sentences. We filter out stories of more than 1,000 words in WritingPrompts, and cut the remaining stories into five-sentence-paragraphs.

For each five-sentence-paragraph, we define the first three sentences as contexts of the base event, the fourth sentence as the base event, and the fifth sentence as the inference target. For example, as shown in Table 3, the first three sentences describe a context that Jason was unsatisfied about his job and applied for a new job. Hence, after happening the event “he got the job”, a plausible react about the event could be “jason was much happier at his new job”. In total, the auxiliary dataset contains 192,316 (*context, event, target*) triples.

4.2 Baselines

We compared our proposed model with the following four baseline methods:

- **RNN-based Seq2Seq** proposed by Sap et al. (2018) for the If-Then reasoning on Atomic.
- **Variational Seq2Seq** combines a latent variable with the encoder-decoder structure through converting the last hidden state of RNN encoder into a Gaussian distributed latent variable (Bowman et al., 2015).
- **VRNMT** Propose by Su et al. (2018), VRNMT combines CVAE with attention-based encoder-decoder framework through introduces a latent variable to model the semantic distribution of targets.
- **CWVAE-Unpretrained** refers to the CWVAE model without the pretrain stage.

Note that, for each baseline method, we train distinct models for each distinct inference dimension, respectively.

4.3 Evaluation Metrics

Automatic Evaluation We first compare the perplexity of CWVAE with baseline methods. Perplexity measures the probability of model to regenerate the exact targets, which is particular suit-

Metric	Methods	xIntent	xReact	oReact
PPL	RNN-based Seq2Seq	44.12	29.18	14.08
	Variational Seq2Seq	42.06	28.22	12.62
	VRNMT	33.45	25.54	11.93
	CWVAE-Unpretrained CWVAE	31.32 29.23	24.07 23.17	11.37 11.04
BLEU	RNN-based Seq2Seq	2.75	2.11	5.18
	Variational Seq2Seq	2.84	2.43	2.08
	VRNMT	4.81	3.94	6.61
	CWVAE-Unpretrained CWVAE	7.36 12.98	5.52 5.65	5.33 6.97

Table 4: Average perplexity and BLEU score (reported in percentages) for the top 10 generations under each inference dimension of Event2Mind. The the best result for each dimension is emboldened.

Metric	Methods	xIntent	xReact	oReact
dist-1	RNN-based Seq2Seq	0.0002	0.0002	0.0001
	Variational Seq2Seq	0.0006	0.0003	0.0001
	VRNMT	0.0002	0.0002	0.0003
	CWVAE-Unpretrained CWVAE	0.0023 0.0052	0.0017 0.0033	0.0004 0.0025
dist-2	RNN-based Seq2Seq	0.0005	0.0002	0.0002
	Variational Seq2Seq	0.0014	0.0002	0.0001
	VRNMT	0.0005	0.0003	0.0001
	CWVAE-Unpretrained CWVAE	0.0061 0.0146	0.0040 0.0099	0.0013 0.0063

Table 5: Distinct-1 and distinct-2 scores for the top 10 generations under each inference dimension of Event2Mind. The the best result for each dimension is emboldened.

able for evaluating the model performance on one-to-many problem (Serban et al., 2017). Further, we employ BLEU score to evaluate the accuracy of generations (Papineni et al., 2002), and the number of distinct n-gram to evaluate the diversity of generations (Li et al., 2016). The distinct is normalized to $[0, 1]$ by dividing the total number of generated tokens.

Human Evaluation Since automatic evaluation of generations is still a challenging task (Liu et al., 2016), we also conduct human evaluations on the model performance. Five human experts are employed to evaluate the *coherence*, *diversity* and *fluency* of generated targets. Experts are asked to vote for if a generation is fluent or coherent for each generated target, and give a 1-5 score for the diversity of generations. For both Event2Mind and Atomic datasets, 100 events are randomly selected from the test set. For each method, top 10 generated targets of each base event are used for eval-

Metric	Methods	xIntent	xNeed	xAttr	xEffect	xReact	xWant	oWant	oReact	oEffect
PPL	RNN-based Seq2Seq	22.54	24.69	33.54	65.13	29.52	26.63	16.76	14.99	35.17
	Variational Seq2Seq	26.48	28.31	33.00	68.62	29.93	29.50	16.98	14.25	34.20
	VRNMT	21.04	24.28	24.87	61.05	26.62	28.57	14.45	14.86	30.12
	CWVAE-Unpretrained	20.73	23.72	25.80	60.62	25.75	26.71	15.93	12.82	32.00
	CWVAE	15.93	20.32	23.85	50.74	21.39	24.02	14.02	11.70	29.13
BLEU	RNN-based Seq2Seq	8.17	12.35	2.96	5.26	3.43	13.44	7.08	4.09	6.42
	Variational Seq2Seq	8.31	12.05	2.13	6.07	2.52	11.71	7.40	4.08	6.38
	VRNMT	9.52	13.35	4.87	4.42	7.64	9.80	10.79	5.28	13.71
	CWVAE-Unpretrained	11.37	14.64	4.07	14.11	7.86	12.70	12.09	8.16	14.93
	CWVAE	12.12	15.67	5.63	14.64	8.13	15.01	13.83	8.58	11.63

Table 6: Average perplexity and BLEU scores (reported in percentages) for the top 10 generations under each inference dimension of Atomic. The the best result for each dimension is emboldened.

Metric	Methods	xIntent	xNeed	xAttr	xEffect	xReact	xWant	oWant	oReact	oEffect
dist-1	RNN-based Seq2Seq	0.0012	0.0029	0.0004	0.0019	0.0001	0.0022	0.0006	0.0001	0.0006
	Variational Seq2Seq	0.0006	0.0018	0.0002	0.0002	0.0001	0.0013	0.0007	0.0001	0.0002
	VRNMT	0.0002	0.0001	0.0053	0.0005	0.0018	0.0022	0.0005	0.0001	0.0004
	CWVAE-Unpretrained	0.0019	0.0036	0.0119	0.0046	0.0021	0.0013	0.0018	0.0005	0.0006
	CWVAE	0.0055	0.0045	0.0142	0.0028	0.0043	0.0040	0.0021	0.0030	0.0033
dist-2	RNN-based Seq2Seq	0.0036	0.0081	0.0002	0.0018	0.0002	0.0006	0.0013	0.0001	0.0011
	Variational Seq2Seq	0.0013	0.0042	0.0001	0.0003	0.0002	0.0026	0.0002	0.0003	0.0006
	VRNMT	0.0002	0.0011	0.0002	0.0005	0.0001	0.0034	0.0005	0.0001	0.0004
	CWVAE-Unpretrained	0.0060	0.0088	0.0136	0.0113	0.0043	0.0029	0.0041	0.0011	0.0009
	CWVAE	0.0162	0.0112	0.0146	0.0072	0.0013	0.0107	0.0044	0.0068	0.0093

Table 7: Distinct-1 and distinct-2 scores for the top 10 generations under each inference dimension of Atomic. The the best result for each dimension is emboldened.

uation. Finally we report three overall averaged scores of *coherence*, *diversity* and *fluency* on both datasets, respectively.

4.4 Overall Results

We list the perplexity and BLEU score of CWVAE and baseline methods on Event2Mind and Atomic in Table 4 and Table 6, respectively, and show the distinct-1 and distinct-2 score on Event2Mind and Atomic in Table 5 and Table 7, respectively. We find that:

(1) As shown in Table 5 and Table 7, comparison between RNN-based Seq2Seq and variational-based methods, including Variational Seq2Seq, VRNMT, CWVAE-unpretrained and CWVAE shows that, variational-based methods could increase the diversity of generations. This confirms one of our motivations that variational-based methods could capture the latent semantic distribution within targets and increase the diversity of If-Then reasoning.

(2) Comparing CWVAE-unpretrained with other baseline methods shows that, in general CWVAE improves the accuracy and diversity on both dataset. These results indicate the efficiency of CWVAE in capturing the latent semantic distribution of targets, and generate more reasonable inferential results.

(3) Comparison between CWVAE and CWVAE-unpretrained shows that the pre-

Methods	Coherence	Diversity	Fluency
RNN-based Seq2Seq	0.28	2.03	0.73
Variational Seq2Seq	0.33	1.67	0.92
VRNMT	0.32	2.60	0.83
CWVAE-Unpretrained	0.36	2.10	0.92
CWVAE	0.43	2.85	0.96

Table 8: Human evaluation results on Event2Mind.

Methods	Coherence	Diversity	Fluency
RNN-based Seq2Seq	0.21	2.66	0.78
Variational Seq2Seq	0.22	2.70	0.90
VRNMT	0.24	2.61	0.78
CWVAE-Unpretrained	0.25	2.72	0.83
CWVAE	0.32	3.03	0.90

Table 9: Human evaluation results on Atomic.

train stage could enhance the performance of CWVAE in both the accuracy and diversity. This is mainly because event knowledge could offer the guidance for If-Then reasoning. In the pretrain stage, CWVAE could capture the event background knowledge through context-aware latent variable, and such knowledge could be adapted to our task through the finetune stage.

To further evaluate the effectiveness of our proposed approach, we also conduct human evaluations, the results of which are shown in Table 8 and Table 9. On both datasets, CWVAE-based methods achieve consistent better coherence, diversity and fluency performances. While comparing with CWVAE-Unpretrained, the pre-train procedure could improve the performance

Base event	Inference dim.	Generations		Ground truth
		CWVAE	RNN-based Seq2Seq	
PersonX works tirelessly	xIntent	be productive and hardworking	get the job done	be productive finish the project as soon as possible reach goal
		finish his work soon	get the job done on time	
		earn more money and accomplish goal	get the job done in the way	
		finish his work and make money	get the job done for his life	
		finish his work and accomplish goal	make money for his own future	
		be productive and successful	make money for his own money	

Table 10: An example of inferences made by CWVAE and RNN-based Seq2Seq model under inference dimension “xIntent”.

on coherence and fluency. The main reasons are twofold: first, the CWVAE has advantage in capturing the semantic distribution of targets; second, event background learned from the pretrain stage is helpful for the If-Then reasoning.

4.5 Case Study

Table 10 provides an example of model generations given the base event “PersonX works tirelessly” and the inference dimension “xIntent”. The generations under CWVAE mainly contain four kinds of semantics: (1) be productive, (2) finish his work soon, (3) accomplish goal, (4) earn more money. While the semantics of generations using baseline RNN-based Seq2Seq model is relatively limited. Furthermore, the first three kinds of semantic overlap the three ground truth targets, and the fourth kind of semantic is in accordance with daily-life commonsense. Compared to RNN-based Seq2Seq model, our approach can increase the diversity and rationality of generations, meanwhile keep the accuracy.

5 Related Work

5.1 Event-Centered Commonsense Reasoning

Understanding events and constructing event-centered commonsense knowledge are crucial to many NLP applications, such as intention recognition (Goldwasser and Zhang, 2016) and dialog generation (Wen et al., 2017). Recently a growing number of studies focus on event-centered commonsense reasoning, which mainly concentrates on two areas, script event prediction and story ending generation/choosing.

Script event prediction concerns with the temporal relationships between script events (Granroth-Wilding and Clark, 2016), which requires models to choose a correct subsequent triple-organized event among the candidates (Wang et al., 2017). Prior work mainly focused on modeling event pairs (Granroth-Wilding and Clark, 2016), event chains (Wang et al., 2017) and event graph (Li et al., 2018) to predict the subse-

quent event. Story ending generation focuses on generating plausible story endings (Mostafazadeh et al., 2016), which requires models to understand the story context, and keep generated endings logically consistent with it (Peng et al., 2017; Guan et al., 2019). The above tasks mainly investigate the logical orders of events, whereas the If-Then reasoning task focuses on inferring the mental state of event participants.

5.2 Variational AutoEncoder-Decoder Based Natural Language Generation

VAE (Kingma and Welling, 2014) has been widely applied in various of text generation tasks, such as dialogue and machine translation.

In dialogue generation, Zhao et al. (2017) adapts VAE with encoder-decoder framework to model the latent semantic distribution of answers, which can increase the diversity of generations. For the task of machine translation, Su et al. (2018) and Zhang et al. (2016) employ a latent variable to capture the semantic interaction between the source and target sentence, and regard the latent variable as a supplementation of attention mechanism. While Wang et al. (2019) use the latent variable to model topic distributions in text generation. In this paper, we introduce an additional context-aware latent variable to effectively learn background knowledge and conduct If-Then reasoning on the guidance of it.

6 Conclusion

In this paper, we propose a novel context-aware VAE (CWVAE) framework with two training stages for If-Then commonsense reasoning. By introducing an additional context-aware latent variable, CWVAE is able to learn external background knowledge, and conduct If-Then reasoning under its guidance. In the pretrain stage, CWVAE learns event background knowledge, then in the finetune stage CWVAE adapts such knowledge to each inference dimension. Experimental results demonstrate that CWVAE outperforms baseline methods in both the accuracy and diversity of generations.

7 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 (SQ2018AAA010010), the National Key Research and Development Program of China (2018YFB1005103), the National Natural Science Foundation of China (NSFC) via Grant 61702137.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.
- Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. *Proc. of ACL-08: HLT*, pages 789–797.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.
- Dan Goldwasser and Xiao Zhang. 2016. Understanding satirical articles using common-sense. *Transactions of the Association for Computational Linguistics*, 4:537–549.
- Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *arXiv preprint arXiv:1709.04348*.
- Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *AAAI*.
- Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge.
- Ting-Hao Kenneth 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.
- D Kinga and J Ba Adam. 2015. A method for stochastic optimization. In *ICLR*, volume 5.
- Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. *stat*, 1050:10.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.
- 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. AAAI Press.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.
- 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 *Proceedings of the 2016 Conference of the NAACL: Human Language Technologies*, pages 839–849.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318.
- Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255.
- Haoruo Peng, Snigdha Chaturvedi, and Dan Roth. 2017. A joint model for semantic sequences: Frames, entities, sentiments. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 173–183.
- Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *Proc. of the 56th ACL*, volume 1, pages 463–473.
- Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2018. Atomic: An atlas of machine commonsense for if-then reasoning. *arXiv preprint arXiv:1811.00146*.
- Roxane Segers, Marco Rospoche, Piek Vossen, Egoitz Laparra, German Rigau, and Anne-Lyse Minard. 2016. The event and implied situation ontology (eso): Application and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

- Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491.
- Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *AAAI*.
- Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, and Lawrence Carin. 2019. Topic-guided variational autoencoders for text generation.
- Zhongqing Wang, Yue Zhang, and Ching-Yun Chang. 2017. Integrating order information and event relation for script event prediction. In *Proc. of the 2017 EMNLP*, pages 57–67.
- Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *Proc. of the 34th ICML-Volume 70*, pages 3732–3741. JMLR. org.
- Biao Zhang, Deyi Xiong, Hong Duan, Min Zhang, et al. 2016. Variational neural machine translation. In *Proc. of the 2016 EMNLP*, pages 521–530.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.