Syntactically Diverse Adversarial Network for Knowledge-Grounded Conversation Generation

Fuwei Cui[1], Hui Di[2], Hongjie Ren[3], Kazushige Ouchi[2], Ze Liu[1] and Jinan Xu[3]*

[1]School of Electronic Information Engineering, Beijing Jiaotong University, China
[2]Toshiba (China) Co., Ltd, China / Japan
[3]School of Computer Information Technology, Beijing Jiaotong University, China
{fuweicui, 20125222, zliu, jaxu}@bjtu.edu.cn
dihui@toshiba.com.cn, kazushige.ouchi@toshiba.co.jp

Abstract

Generative conversation systems tend to produce meaningless and generic responses, which significantly reduce the user experience. In order to generate informative and diverse responses, recent studies proposed to fuse knowledge to improve informativeness and adopt latent variables to enhance the diversity. However, utilizing latent variables will lead to the inaccuracy of knowledge in the responses, and the dissemination of wrong knowledge will mislead the communicators. To address this problem, we propose a Syntactically Diverse Adversarial Network (SDAN) for knowledge-grounded conversation model. SDAN contains an adversarial hierarchical semantic network to keep the semantic coherence, a knowledge-aware network to attend more related knowledge for improving the informativeness and a syntactic latent variable network to generate syntactically diverse responses. Additionally, in order to increase the controllability of syntax, we adopt adversarial learning to decouple semantic and syntactic representations. Experimental results show that our model can not only generate syntactically diverse and knowledge-accurate responses but also significantly achieve the balance between improving the syntactic diversity and maintaining the knowledge accuracy.

1 Introduction

Nowadays, conversation generation has become a research hotspot because of its wide application, such as voice assistant, customer service assistant and chat robot (Cui et al., 2021). The goal of conversation model is to generate diverse and informative responses like human. Although the existing models have achieved promising performance, they still suffer from generating general and meaningless responses (Wu et al., 2020), which significantly disrupt the user experience. Consequently, it is very crucial and urgent to generate high-quality responses.

To generate high-quality responses, many researches have been proposed to improve informativeness or diversity of responses. For informative responses, some early studies utilize context information to the decoding process (Sordoni et al., 2015; Yao et al., 2015). After that, researchers extract topic information from context (Hedayatnia et al., 2020) or add external topic to the decoder (Xing et al., 2016, 2017). Lately, researchers focus on fusing knowledge into conversation model (Ghazvininejad et al., 2018; Zhou et al., 2018; Lian et al., 2019; Wu et al., 2020; Lin et al., 2020). Although the knowledge-grounded model can generate informative responses with accurate knowledge, which may generate responses that lack diversity. For diverse responses, previous studies generally adopt beam search algorithm (Li et al., 2016b) and its variants to improve diversity (Vijayakumar et al., 2016). In recent year, latent variables are widely used in conversation model, and can significantly enhance the diversity (Serban et al., 2017; Zhao et al., 2017; Park et al., 2018; Shen et al., 2019; Ruan et al., 2019; Cui et al., 2021), and generative adversarial networks (GAN) (Xu et al., 2018) and reinforcement learning (RL) (Sankar and Ravi, 2019) are also adopted to generate diverse responses. Although the introduction of hidden variables can increase diversity while maintaining semantic consistency, it may lead to inaccuracy in decoding specific knowledge, because the latent variables may generate semantically similar responses with a certain probability. For example, as shown in Table 1, there is a song name (*Be Your Girl All Your Life*) in query, where the response R1 will be generated by the variational latent model. In R1, the song name may be decoded as *Be Your Woman in The Next Life*, which is another song name. Then, the wrong responses will be generated. How to improve diversity of responses and

*Jinan Xu is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4620–4630
November 7–11, 2021.

Example	Knowledge Triple		
	Head Entity	Relation	Tail Entity
Query: Who is singer of *Be Your Girl All Your Life*? 《一辈子做你的女孩》的演唱者是谁?	《一辈子做你的女孩》 *Be Your Girl All Your Life*	演唱者 singer	萧亚轩 Elva Hsiao
Golden Response: The singer of *Be Your Girl All Your Life* is Elva Hsiao. 《一辈子做你的女孩》的演唱者是萧亚轩。 √			
Response1: The singer of *Be Your Woman in The Next Life* is Meizi Long. 《下辈子做你的女人》的演唱者是龙梅子。 ×			
Response2: Elva Hsiao sang *Be Your Girl All Your Life*. 萧亚轩演唱了《一辈子做你的女孩》。 √			

Table 1: An illustrative example. Response1 shows the response generated with semantic latent variable, Response2 shows the response generated with syntactic latent variable. √ and × denote that the responses are right and wrong, respectively.

preserve the accuracy of knowledge simultaneously is a huge challenge in knowledge-grounded conversation generation.

To tackle this challenge, we propose a Syntactically Diverse Adversarial Network (SDAN) for knowledge-grounded conversation generation. First, we utilize a hierarchical network to model the semantic information of context and an adversarial network to prevent semantic information from affecting syntactic information. Next, we adopt a knowledge-aware network to represent the knowledge related to the query, which takes attention mechanism to capture more important knowledge. Then, we design a syntax encoder to model syntax information and use a latent variable to keep the syntactic diversity. Finally, the encoded knowledge, syntax and context are concatenated together to initialize the decoder. Additionally, we employ adversarial network to keep the separation of syntax and semantics to prevent their mutual influence. The results of experiments on KdConv datasets show that our model can achieve better trade-off between improving diversity and maintaining knowledge accuracy than baselines.

Our main contributions are as follows:

- To best of our knowledge, we are the first to adopt syntactic latent variable to simultaneously improve the diversity and maintain the accuracy of knowledge in knowledge-grounded conversation generation, and propose a novel Syntactically Diverse Adversarial Network.
- Our model gains competitive diversity scores and the best knowledge-accurate scores than baselines.
- We further conduct extensive ablation studies on the proposed several components. These analyses explore intuitive interpretability of why do the adversarial network, knowledge and syntactic latent variable have an effect on our model, and provide a reference for future model design.

2 Background

2.1 Variational Autoencoder

Since our model adopts latent variables, we briefly review the architecture of Variational Autoencoder (VAE) (Kingma and Welling, 2014), a generative model which utilizes a latent variable z to encode the information of the utterance x, and then decodes the original x from z. The probability of x can be computed as follows:

$$p(x) = \int p(x,z)d_z = \int p(z)p(x|z)d_z \quad (1)$$

where $p(z)$ is the prior distribution, $p(x|z)$ is given by the decoder. Since the integral is unavailable in closed form (Blei et al., 2017), the VAE is trained by maximizing the *evidence lower bound* (ELBO), which is defined as follows:

$$\begin{aligned} &logp(x) \geq \text{ELBO} \\ &= \mathop{\mathbb{E}}_{q(z|x)}[logp(x|z)] - D_{KL}(q(z|x)||p(x)) \end{aligned} \quad (2)$$

where $q(z|x)$ is posterior distribution obtained by the encoder, $\mathbb{E}$ is mathematical expectation, $D_{KL}(\cdot||\cdot)$ indicates the Kullback-Leibler(KL) Divergence which is utilized to represent the similarity of two distributions.

2.2 Generated Adversarial Learning

Generated Adversarial Learning (GAN) (Goodfellow et al., 2014) is widely used in the generation of image and text, which consists of a Generator (G)

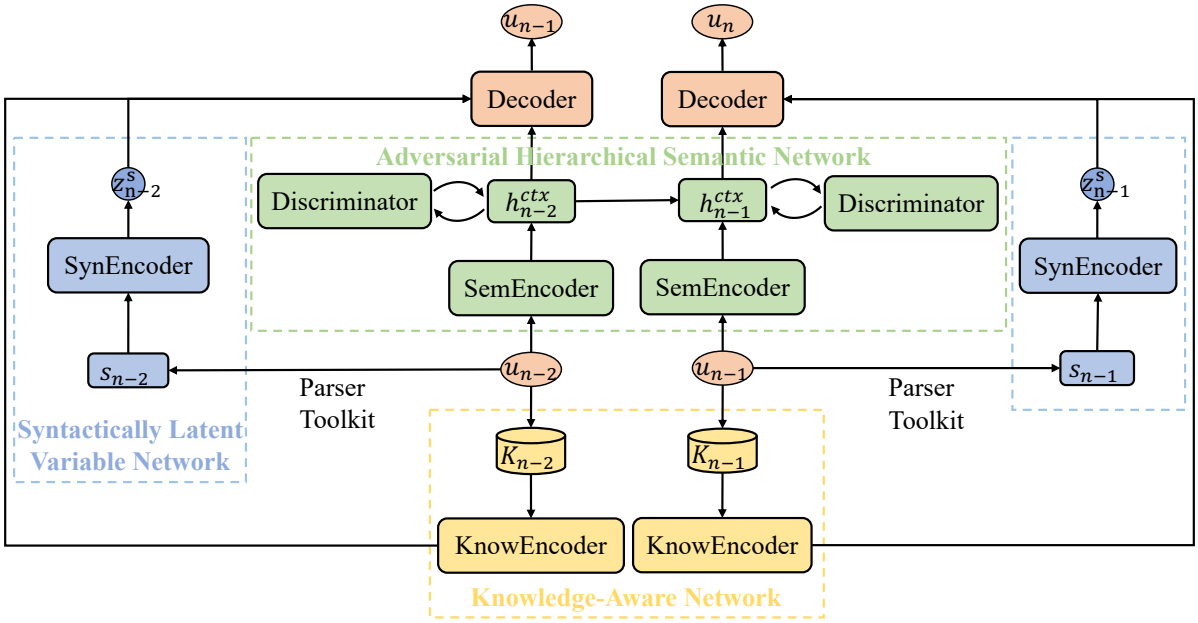

Figure 1: Overview of SDAN, combining a adversarial hierarchical semantic network to model the semantics, a knowledge-aware network to represent knowledge and a syntactically latent variable network to control the diversity of syntax. u_i denotes the i-th utterance. h_i^{ctx} represents the context information. K_i is the relevant knowledge. s_i is the syntax of u_i obtained by the Parser Toolkit. z_i^s denotes the syntactic latent variable. The more details of SDAN are shown in Section 3.

and a Discriminator (D). The training objective of GAN is defined as follows:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim P_z(z)}[log(1-D(z))] \quad (3)$$

where G is utilized to obtain the generated distribution $p_g(x)$ from noisy distribution $p_z(z)$ to approximate the true distribution $p_{data}(x)$, and D is used to distinguish the distribution of $p_g(x)$ and $p_{data}(x)$. G attends to reduce the value of V to make the generated distribution unrecognized, but D intends to enlarge the value of V to effectively identify the true and false classes of data. In the process of training, G and D are optimized alternately, and the optimal solution can be achieved by iterating for many times.

3 Methodology

3.1 Task Formulation and Model Overview

Formally, we assume the training data $\mathcal{D}$ consists of N samples of conversations$\{c_1, c_2, ..., c_N\}$ where each c_i is a sequence of utterances $\{u_1, u_2, ..., u_n\}$ which is expressed as $\{u_t\}_{t=1}^n$. We consider the $\{u_t\}_{t=1}^{n-1}$ as query, the $\{u_t\}_{t=2}^n$ as response. Each query has m related knowledge $(k_1, \ldots, k_m)$, where each knowledge k_i is a triplet (h_i, r_i, t_i), and h_i, r_i and t_i are the head entity, the relation and the tail entity, respectively. Each utterance has the syntax s_i. The goal of our method is to generate informative and diverse responses, so we will fuse knowledge and syntax to the generative model.

The overview of SDAN is shown in Figure 1. The **Adversarial Hierarchical Semantic Network** consists of encoder layer and context layer, which is utilized to model the semantic information. The **Knowledge-Aware Network** adopts attention mechanism to focus the more important knowledge. The **Syntactically Latent Variable Network** adopts a latent variable to generate responses with diverse syntax. Finally, the semantic information, knowledge and syntax from above three networks are concatenated together to the **Decoder**.

3.2 Adversarial Hierarchical Semantic Network

The Hierarchical Semantic Network consists of two layer neural networks. Each input utterance u_i is encoded into a vector h_t^{enc} by the encoder RNN, which is shown as follows:

$$h_t^{enc} = f_\theta^{enc}(u_t) \quad t = 1, \ldots, n \quad (4)$$

where $f_\theta^{enc}(\cdot)$ is a bidirectional gated recurrent unit (BiGRU).

The context vector h_t^{ctx} represents the historical information, which updates its hidden states by using the encoder vector h_t^{enc} and is calculated by:

$$h_t^{ctx} = f_\theta^{ctx}(h_{t-1}^{ctx}, h_t^{enc}) \tag{5}$$

where the initial value of h_t^{ctx} is 0.

The semantic information from the hierarchical semantic network may contain the syntactic information, which can lead to poor syntactic controllability. In order to solve this problem, we introduce adversarial network to prevent semantic information from containing syntactic information. Specifically, we introduce a discriminator to predict the syntax tree sequence s_t according to the semantic information of the context h_t^{ctx}. The context layer and encoder layer can be regarded as the generator. The generator is trained to learn the semantic information to prevent the discriminator predicting the syntax from the semantic information and to cheat the discriminator by maximizing the adversarial loss, that is, minimizing the following formula:

$$loss_{syn}^{adv} = \sum_{t=1}^{t=n} logp_{adv}(s_t|h_t^{ctx}) \tag{6}$$

3.3 Knowledge-Aware Network

The knowledge can be retrieved from the knowledge base to select the related knowledge. The knowledge used in this paper is given in the dataset and one query may have multiple knowledge, so we employ attention mechanism to pay more attention to the important knowledge, which is similar to (Zhou et al., 2020).

We assume that there is m related knowledge $(k_1, \ldots, k_m)$ given for a query u_t, and each knowledge k_i is a triplet (h_i, r_i, t_i). First, we treat the average word embeddings of h_i and r_i as the key vector $kv_i (i = 1, \ldots, m)$. Then, we use the word embedding of the query u to attend to kv_i:

$$\alpha_i = \mathrm{softmax_i}(\mathrm{emb}(u_t)^T kv_i) \tag{7}$$

where $\mathrm{emb}(\cdot)$ is the embedding vector, $\mathrm{softmax}(\cdot)$ is a generalization of the logistic function which normalizes all values between 0 and 1. After that, we obtain the knowledge k_t by summing all the weighted tail entity t_i:

$$k_t = \sum_{i=1}^{i=m} \alpha_i t_i \tag{8}$$

Finally, we utilize a BiGRU to encode the knowledge to model the knowledge vector h_t^{kno}, which is computed as follows:

$$h_t^{kno} = f_\theta^{kno}(k_t) \tag{9}$$

where $f_\theta^{kno}(\cdot)$ is a BiGRU.

3.4 Syntactically Latent Variable Network

Each utterance contains syntactic information, which is usually represented by syntactic tree. The syntactic tree can be modeled by a neural network or obtained by the parser toolkit. In this paper, we first utilize the Stanford Parser toolkit [1] to process all the utterances in the dataset to get their syntactic tree sequences, which contain the syntactic tokens and the brackets (the brackets represent the syntactic structures). Then, a SynEncoder is employed to represent the syntactic vector h_t^{syn}, which is shown as follows:

$$h_t^{syn} = f_\theta^{syn}(s_t) \tag{10}$$

where $f_\theta^{syn}(\cdot)$ is a BiGRU, s_t is the syntactic tree sequence.

Finally, in order to generate syntactically diverse responses, we adopt a syntactic latent variable z_t^s to control the syntactic information. We define the prior distribution of z_t^s as:

$$p_\theta(z_t^s|s_t) = \mathcal{N}(z|\mu^s, {\sigma^s}^2 I) \tag{11}$$

where $\mathcal{N}(\cdot)$ is a Gaussian distribution, μ^s and σ^s are the means and the diagonal variances of the prior distributions, respectively, which are calculated as:

$$\mu^s = \mathrm{MLP}_\theta(h_t^{syn}) \tag{12}$$

$$\sigma^s = \mathrm{Softplus}(\mu^s) \tag{13}$$

where $\mathrm{MLP}_\theta(\cdot)$ is a feed-forward neural network and $\mathrm{Softplus}(\cdot)$ is an activation function which can keep the result positive.

For the posterior distribution of z_t^s, we use h_t^s and h_{t+1}^s to calculate it in training set (h_t^s in test set):

$$q_\phi(z_t^s|s_t, s_{t+1}) = \mathcal{N}(z|\mu^{s'}, {\sigma^{s'}}^2 I) \tag{14}$$

where

$$\mu^{s'} = \mathrm{MLP}_\phi(h_t^{syn}, h_{t+1}^{syn}) \tag{15}$$

$$\sigma^{s'} = \mathrm{Softplus}(\mu^{s'}) \tag{16}$$

[1] https://nlp.stanford.edu/software/lex-parser.shtml

3.5 Decoder

From the three networks mentioned above, we obtain the representation of semantics, knowledge and syntax, We concatenate them together to be the initial state of the decoder,which is shown as follows:

$$h_{ini}^{dec} = [h_t^{ctx}, h_t^{kno}, z_t^s] \quad (17)$$

Finally, we output the response u_{t+1}, which is shown as follows:

$$u_{t+1} = f_\theta^{dec}(h_{ini}^{dec}) \quad (18)$$

where $f_\theta^{dec}(\cdot)$ is a GRU.

3.6 Training Objective

Because of the existence of latent variables in our model, the training objective of latent variables is to maximize the following ELBO:

$$\begin{aligned} &\text{ELBO} = loss_{rec} + loss_{KL} \\ &- \mathbb{E}_{q_\phi}[log_{p_\theta}(u_n|\{u_t\}_{t=1}^{n-1}, \{k_t\}_{t=1}^{n-1}, \{s_t\}_{t=1}^{n-1})] \\ &+ D_{KL}(q_\phi(z^s|\{s_t\}_{t=1}^{n})||p_\theta(z^s|\{s_t\}_{t=1}^{n-1})) \end{aligned} \quad (19)$$

where $loss_{rec}$ is the reconstruction loss, $loss_{KL}$ is the KL divergence to represent the similarity of the posterior distribution and the prior distribution of the latent variable z_t^s.

Then, the final objective is to minimize the following formula:

$$min[loss_{syn}^{adv}] + min[-\text{ELBO}] \quad (20)$$

where the two losses are optimized iteratively.

4 Experiments

4.1 Experiment Setting

Datasets: We conduct our experiments on **KdConv** (Zhou et al., 2020) dataset, which is a Chinese multi-domain knowledge-driven conversation dataset. This dataset contains 4.5K conversations from three domains (film, music, and travel), and 86K utterances with an average turn number of 19.0. In **KdConv**, each utterance has 0 to m pieces of knowledge, and the value of m is different for each utterance.

Hyper-parameters: In our model, we employ GRU as our base cell. The dimension of embedding, hidden layer and latent variable layer are set to 500, 1000 and 100, respectively. We use Adam (Kingma and Ba, 2015) as our optimizer. The max length of sentences is set to 20. The learning rates of generator and discriminator are 1e-4 and 1e-5, respectively. The mini-batch size is set to 32. In order to avoid the notorious degeneration problem (Bowman et al., 2016; Chen et al., 2017), we employ KL annealing, and the step of which is set to 25000.

Baseline Models: We compare our model with two baselines. They all focus on knowledge-grounded multi-turn conversation: 1) Hierarchical Recurrent Encoder-Decoder (HRED) (Serban et al., 2016) + knowledge (Zhou et al., 2020); 2) Variational Hierarchical Recurrent Encoder-Decoder (VHRED) (Serban et al., 2017) with KL annealing + knowledge.

4.2 Evaluation Design

We evaluate the generated responses from two aspects: automatic evaluation metrics and manual evaluation metrics.

For automatic evaluation metrics, we utilize four classes of evaluation metrics: **Token-level Metrics**: Perplexity (PPL) is used to evaluate whether the generated response is grammatical and fluent. **Overlapping-based Metrics**: We adopt the BLEU-2/3 (Papineni et al., 2002) to evaluate the reconstruction performance, which can reflect how well the model could preserve information from knowledge and ground truth response. **Embedding-based Metrics**: Average, Greedy and Extrema are adopted to measure the semantic similarity between words in generated response and the ground truth. **Diversity**: We employ Dist-1/2 (Li et al., 2016a) to measure the diversity of the responses, which are defined as the ratio of distinct uni/bi-grams. **Knowledge Utilization**: E_{match} is the averaged number of the entities matched with the related knowledge triplets in the responses (Zhou et al., 2020; Wu et al., 2020).

For manual evaluation metrics, three evaluation metrics are adopted, which range from 1 to 5:

Coherence (Cohe) denotes the semantic similarity of response and query: ① score 1: The response and query are completely different and semantically different. ② score 2: The response and query are completely different, but a little semantically similar. ③ score 3: The response and query are partly the same, but semantically similar. ④ score 4: The response and query are mostly the same, but semantically very similar. ⑤ score 5: The response and query are exactly the same.

Model	Average	Extrema	Greedy	BLEU-2/3	Dist-1/2	PPL	E_{match}
Film							
HRED+know	**0.842**	**0.638**	0.681	9.454 / 5.503	0.238 / 0.488	27.950	0.95
VHRED+know	0.840	0.634	**0.690**	7.276 / 3.575	**0.319 / 0.717**	**14.258**	0.86
SDAN(Ours)	0.838	0.637	0.683	**9.614 / 5.795**	0.243 / 0.551	18.714	**1.02**
Music							
HRED+know	0.840	0.645	0.714	14.653 / 9.799	0.274 / 0.567	25.359	0.98
VHRED+know	**0.846**	0.646	0.713	12.837 / 8.534	**0.310 / 0.682**	**11.672**	0.90
SDAN(Ours)	0.843	**0.648**	**0.718**	**14.882 / 10.150**	0.288 / 0.595	15.006	**1.21**
Travel							
HRED+know	**0.858**	0.684	0.761	**22.499** / 18.001	0.268 / 0.547	10.604	1.21
VHRED+know	0.854	0.681	0.751	20.463 / 15.969	**0.301 / 0.652**	**6.528**	0.97
SDAN(Ours)	0.852	**0.689**	**0.761**	22.451 / **18.030**	0.276 / 0.571	8.069	**1.25**

Table 2: Automatic evaluation results on KdConv Corpus. The best results are in **bold**. The "+know" means the models are enhanced by the knowledge base. The VHRED+know and SDAN have the semantic and syntactic latent variables, respectively.

Fluency (Flu) represents the grammatical problem: ① score1: The response can not understand. ② score2: The response has more than four grammatical errors and is difficult to understand. ③ score3: The response has three or four grammatical errors and is not fluent. ④ score4: The response has one or two grammatical errors and is fluent. ⑤ score5: The response has no grammatical errors and is fluent.

Informativeness (Info) is designed to measure whether the response is relevant to the knowledge information: ① score 1: The response does not contain the relevant knowledge and relevant to the context. ② score 2: The response does not contain the relevant knowledge, but relevant to the context. ③ score 3: The response only contains one relevant knowledge. ④ score 4: The response contains part of the relevant knowledge. ⑤ score 5: The response contains all the relevant knowledge.

Model	Cohe	Flu	Info
Film $\setminus \kappa$	0.62	0.53	0.75
HRED+know	2.12	2.65	2.24
VHRED+know	2.20	**3.03**	1.97
SDAN(Ours)	**2.25**	2.86	**2.27**
Music $\setminus \kappa$	0.6	0.43	0.71
HRED+know	**2.32**	2.91	2.35
VHRED+know	2.27	**3.25**	2.06
SDAN(Ours)	2.30	3.03	**2.41**
Travel $\setminus \kappa$	0.78	0.58	0.83
HRED+know	2.48	3.24	2.43
VHRED+know	2.51	**3.51**	2.11
SDAN(Ours)	**2.55**	3.42	**2.47**

Table 3: Manual evaluation results on Kdconv Corpus. κ is the Fleiss' kappa value.

4.3 Results of Automatic Evaluation

The results of automatic evaluation metrics are shown in Table 2. We analyze the results from the following perspective:

The influence of semantic and syntactic latent variables:

1) Although our improvement on some domains is limited, but we achieve balance between syntactic diversity and knowledge accuracy.

2) In terms of embedding-based metrics (Average, Extrema and Greedy), there is little difference among the three models. So we can conclude that adopting the semantic and syntactic latent variables have little effect on the semantics of responses.

3) Compared with HRED+know, VHRED +know obtains lower BLEU-k scores and higher Dist-k scores, and SDAN performs better in these two aspects. We can find that although semantic hidden variables can significantly improve the diversity, but also greatly reduce the accuracy of responses. But the syntactic latent variables can not only improve the diversity but also enhance the accuracy of responses. The reason is that semantic latent variables may utilize other words with similar semantics, which will lead to the inaccuracy of the knowledge, while the syntactic latent variables only change the syntax of responses, which has no influence on the accuracy of knowledge.

Model	Average	Extrema	Greedy	BLEU-2/3	Dist-1/2	PPL	E_{match}
Film							
SDAN(Ours)	0.838	0.637	**0.683**	**9.614 / 5.795**	**0.243 / 0.551**	18.714	**1.02**
-adv	0.828	0.615	0.658	9.106 / 5.491	0.240 / 0.529	**17.848**	0.97
-adv-know	0.774	0.546	0.575	4.145 / 2.275	0.109 / 0.231	20.551	0.51
-adv-syn	**0.842**	**0.638**	0.681	9.454 / 5.503	0.238 / 0.488	27.950	0.99
-adv-know-syn	0.814	0.587	0.635	4.491 / 2.315	0.031 / 0.044	22.615	0.56
Music							
SDAN(Ours)	**0.843**	**0.648**	**0.718**	**14.882 / 10.150**	**0.288 / 0.595**	**15.006**	**1.21**
-adv	0.838	0.635	0.704	14.082 / 9.150	0.276 / 0.575	15.161	0.16
-adv-know	0.811	0.577	0.611	4.623 / 2.440	0.126 / 0.234	19.632	0.53
-adv-syn	0.840	0.645	0.714	14.653 / 9.799	0.274 / 0.567	25.359	1.19
-adv-know-syn	0.794	0.548	0.591	4.754 / 2.472	0.026 / 0.034	19.818	0.59
Travel							
SDAN(Ours)	0.852	**0.689**	**0.761**	22.451 / **18.030**	**0.276 / 0.571**	**8.069**	**1.25**
-adv	0.832	0.639	0.711	21.987 / 17.993	0.256 / 0.531	8.148	1.21
-adv-know	0.766	0.547	0.575	3.772 / 1.935	0.148 / 0.262	11.320	0.54
-adv-syn	**0.858**	0.684	0.761	**22.499** / 18.001	0.268 / 0.547	10.604	1.25
-adv-know-syn	0.747	0.500	0.595	3.561 / 1.935	0.053 / 0.061	11.176	0.58

Table 4: Ablation study on KdConv Corpus. The "-adv", "-know" and "-syn" mean that we eliminate the adversarial network (discriminator), knowledge-aware network and syntactically latent variable network, respectively.

4) It can be seen that the Dist-k scores of VHRED+know is higher than SDAN, which indicates that semantic latent variables are more effective than syntactic latent variables in improving diversity. The reason may be that the vocabularies of semantics are much larger than syntactic vocabularies.

5) For PPL, VHRED+know obtains the best results and SDAN performs better than HRED+know, which denotes that both of the semantic and syntactic latent variables have the positive influence on generating fluent responses and the former works better.

Comparison between domains:

The performance on BLEU-k improves from film domain to travel domain, because there are 1,837 entities and 318 relations in the film domain and 699 entities and 7 relations in the travel domains. The more diverse knowledge increases the difficulty of knowledge selection for knowledge-aware network.

4.4 Results of Manual Evaluation

The results of manual evaluation metrics are shown in Table 3. The scores of three evaluation metrics range from 1 to 5. Additionally, we choose 3 annotators to evaluate the responses generated by the above models, and randomly select 50 conversations from the test set.

For Coherence, the three models are similar in maintaining semantic consistency, which agrees with the results of automatic evaluation. VHRED+know achieves the best Fluency scores and the worst Informativeness scores, which proves that the semantic latent variable can lead to the inaccuracy of knowledge, but can improve the fluency of responses again. Our model obtains the competitive Coherence, Fluency scores and the best Informativeness scores, which indicate that our model can not only generate informative responses but also keep the semantic coherence.

4.5 Ablation Study

To analyze which components are driving the improvements, we present an ablation study in Table 4. We eliminate the adversarial network (discriminator), knowledge-aware network and syntactically latent variable network one by one, which result in four models. The four models are represented as "-adv", "-adv-know", "-adv-syn" and "-adv-know-syn" respectively. By comparing the four models with our SDAN, we can make some conclusions as follows:

1) After eliminating the adversarial network (comparing SDAN with "-adv"), "-adv" has worse performance than SDAN, which indicates that the

Conversation (Film)	Knowledge Triple		
	Head Entity	Relation	Tail Entity
User1: 看过《完美的世界》吗? Have you seen *A Perfect World*? **User2:** 看过，这是一部很好看的惊悚片。 Yes, it's a good thriller. **User1:** 嗯，我也喜欢看这类的电影，138分钟的片长有点受折磨啊，哈哈。 Well, I also like watching this kind of movies. The 138 minute film is a bit tortured. Ha ha.	《完美世界》 *A Perfect World*	类型 Type	惊悚片 Thriller
User2: 哈哈，克林特·伊斯特伍德执导的电影可是很细致的哦。 Ha ha, the film directed by Clint Eastwood is very detailed. **HRED:** 是的，他是一名演员，导演和编剧。你知道他是哪年出生的吗? Yes, he is an actor, director and screenwriter. Do you know when he was born? **HRED+know:** 是的，这部电影还是由克林特·伊斯特伍德制作的呢，知道编剧是谁吗? Yes, the film is still made by Clint Eastwood. Do you know who the screenwriter is? **VHRED+know:** 还是一部由美国制片的呢，你知道主演都有谁吗? It's a movie made in America. Do you know who the stars are? **SDAN(Ours):** 我知道，是美国的克林特·伊斯特伍德制片的。 你知道这部电影的票房怎么样吗? I know, it's made by Clint Eastwood in the United States. Do you know what the box office of the film is like?		电影时长 Film length	138分钟 / 125分钟（德国电视版） 138 minutes / 125 minutes (German TV Version)
		导演 Director	克林特·伊斯特伍德 Clint Eastwood

Table 5: Examples generated by HERD, all baselines and our SDAN from film domain.

adversarial network is effective to enhance the semantics, knowledge accuracy, distinct and fluency, and it is necessary to decouple semantics from syntax.

2) When further removing the knowledge-aware network (comparing "-adv" with "-adv-know"), all the results are worse again, especially the decline of BLEU-k scores is obvious, which denotes that introducing knowledge is essential for conversation generation.

3) While eliminating the syntactically latent variable (comparing "-adv" with "-adv-syn" or comparing "-adv-know" with "-adv-know-syn"), it can be seen that there is a slight improvement in the scores of Average, Extrema, Greedy and BLEU-k, and a bit of lower in the scores of Dist-k, which prove that adopting syntactically latent variable can slightly reduce the semantic consistency and knowledge accuracy, but improve the diversity. Moreover, when the syntactic information and semantic representation exist simultaneously, it certainly need to decouple them by utilizing adversarial network to prevent the influence between them.

4.6 Case Study

The generated responses of HRED, all baselines and our model sampled from test set in film domain are shown in Table 5. As it can be seen, HRED tends to generate generic or irrelevant responses. After introducing knowledge, HRED+konw can generate coherent and informative responses related to the given knowledge. When adopting semantic latent variable, VHRED+know prefer generating responses relevant to the context. while utilizing knowledge and syntactically latent variable, our model can generate knowledge-coherent and diverse responses.

5 Related Work

Sequence-to-sequence (Seq2Seq) model (Sutskever et al., 2014; Shang et al., 2015) with attention (Bahdanau et al., 2015; Cho et al., 2015) has been widely used in the conversation generation. However, models tend to generate meaningless and generic responses (Serban et al., 2017). To alleviate this issue, researchers have utilized context (Sordoni et al., 2015; Yao et al., 2015), topic information (Xing et al., 2016, 2017; Wu et al., 2019) or knowledge (Lian et al., 2019; Wu et al., 2020; Lin et al., 2020) to enhance response quality. The studies of knowledge-grounded conversation generation mainly focus on the method of knowledge retrieval (Lian et al., 2019) or knowledge fusion (Wu et al., 2020; Lin et al., 2020; Ye et al., 2020; Liang et al., 2021) with copy mechanism. The knowledge-grounded models can improve the ac-

curacy of knowledge, but the responses generated by some of them may lack the diversity, which is also a significant reason for generating generic responses.

Recently, to tackle the lack of diversity, researchers have begun to introduce the beam search algorithm (Li et al., 2016b; Vijayakumar et al., 2016) to decoder or latent variables (Serban et al., 2017; Park et al., 2018; Shen et al., 2019). Adopting latent variables can significantly improve the diversity of responses, but it will lead to the inaccuracy of knowledge. To the best of our knowledge, this problem has not been investigated in conversation generation so far.

Different from all the models mentioned above, our approach introduces syntax to conversation generation. We propose a syntactically diverse adversarial network, which utilizes latent variables to control the syntactic diversity. Additionally, we utilize adversarial learning to preserve the disentanglement of syntax and semantics for preventing them from influencing each other. Our model can not only generate sentences with diverse syntax but also keep the accuracy of knowledge.

6 Conclusion

In this paper, we propose a Syntactically Diverse Adversarial Network for knowledge-grounded conversation model, which utilizes adversarial hierarchical semantic network, knowledge-aware network and syntactical latent variable network to model the semantics, knowledge and diverse syntax information. Moreover, our model adopts adversarial learning to enhance the controllability of syntax. According to automatic and manual evaluation, our model competitively improves the quality of generated responses, and obtains better trade-off between improving the diversity and preserving the knowledge accuracy.

Acknowledgements

The research work descried in this paper has been supported by the National Key R&D Program of China (2019YFB1405200) and the National Natural Science Foundation of China (No. 61976015, 61976016, 61876198 and 61370130). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2017. Variational lossy autoencoder. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

K. Cho, A. Courville, and Y. Bengio. 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11):1875–1886.

F. Cui, Q. Cui, and Y. Song. 2021. A survey on learning-based approaches for modeling and classification of human–machine dialog systems. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1418–1432.

Fuwei Cui, Hui Di, Lei Shen, Kazushige Ouchi, Ze Liu, and Jinan Xu. 2021. Modeling semantic and emotional relationship in multi-turn emotional conversations using multi-task learning. *Applied Intelligence*, pages 1573–7497.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

Behnam Hedayatnia, Seokhwan Kim, Yang Liu, Karthik Gopalakrishnan, Mihail Eric, and Dilek Hakkani-Tür. 2020. Policy-driven neural response generation for knowledge-grounded dialogue systems. *CoRR*, abs/2005.12529.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5081–5087. ijcai.org.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13343–13352. AAAI Press.

Xiexiong Lin, Weiyu Jian, Jianshan He, Taifeng Wang, and Wei Chu. 2020. Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 41–52. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1792–1801. Association for Computational Linguistics.

Yu-Ping Ruan, Zhen-Hua Ling, Quan Liu, Jia-Chen Gu, and Xiaodan Zhu. 2019. Promoting diversity for end-to-end conversation response generation. *CoRR*, abs/1901.09444.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 1–10. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *CoRR*, abs/1503.02364.

Lei Shen, Yang Feng, and Haolan Zhan. 2019. Modeling semantic relationship in multi-turn conversations with hierarchical latent variables. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5497–5502. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'14*, page 3104–3112, Cambridge, MA, USA. MIT Press.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5811–5820. Association for Computational Linguistics.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2016. Topic augmented neural response generation with a joint attention mechanism. *CoRR*, abs/1606.08340.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*, pages 3351–3357.

Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. 2018. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, Brussels, Belgium. Association for Computational Linguistics.

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model. *CoRR*, abs/1510.08565.

Hao-Tong Ye, Kai-Lin Lo, Shang-Yu Su, and Yun-Nung Chen. 2020. Knowledge-grounded response generation with deep attentional latent-variable model. *Computer Speech Language*, 63:101069.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7098–7108. Association for Computational Linguistics.