Journal of Computers Vol. 32 No. 2, 2021, pp. 61-73
doi:10.3966/199115992021043202006

Dirichlet Variational Autoencoder for Joint Slot Filling and Intent Detection

Wang Gao[1*], Yu-Wei Wang[1], Fan Zhang[2], Yuan Fang[3]

[1] School of Artificial Intelligence, Jianghan University, Wuhan 430056, China
{gaowang2000, wangyuwei}@foxmail.com

[2] College of Computer Science, Wuhan Donghu University, Wuhan 430212, China
whzhangfan@126.com

[3] School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China
fangyuan2000@foxmail.com

Received 5 March 2020; Revised 12 August 2020; Accepted 15 September 2020

Abstract. Spoken Language Understanding (SLU) is an important part of spoken dialogue systems, which involves two subtasks: slot filling and intent detection. In the SLU task, joint learning has proven effective because intent classes and slot labels can share semantic information with each other. However, because of the high cost of building manually labeled datasets, data scarcity has become a major bottleneck for domain adaptation in SLU. Recent studies on text generation models, such as Dirichlet variational autoencoders (DVAE), have shown excellent results in generating natural sentences and semi-supervised learning. Inspire by this, we first propose a new generative model DVAE-SLU that exploits DVAE's generative ability to generate complete labeled utterances. Furthermore, based on DVAE-SLU, we propose a semi-supervised learning model SDVAE-SLU for joint slot filling and intent detection. Unlike previous methods, this is the first work to generate SLU datasets using DVAE. Experimental results on two classic datasets demonstrate that compared with baseline methods, existing SLU models achieve better performance by training synthetic utterances generated by DVAE-SLU, and the effectiveness of SDVAE-SLU.

Keywords: Dirichlet variational autoencoder, spoken language understanding, semi-supervised learning, data augmentation

1 Introduction

Establishing an intelligent human-machine dialogue system is an important goal in the field of Artificial Intelligence (AI). The system is able to understand human language and give fluent and correct responses. A classic dialogue system contains the following components: (1) Automatic speech recognition converts human speech into corresponding texts. (2) Spoken language understanding (SLU) extracts the semantic representation of the text. (3) Dialogue manager determines the best system response action based on the semantic information [1].

In this paper, we focus on SLU that is the second component of spoken conversation systems. Scarcity of corpus resources has long been plaguing many natural language understanding (NLP) tasks such as SLU. The amount of SLU data for training is limited because the cost of creating manually labeled datasets is quite high, and the domains that require new annotated labeled datasets are almost unlimited [2]. In addition, domains with existing datasets may also experience severe data sparsity issues. The reason is that most SLU datasets are small in size, making it difficult to cover all possible label pairs.

In recent years, significant progress has been made in variational autoencoders (VAE) [3] for text generation. Many studies have improved model performance by generating data augmentation and semi-

* Corresponding Author

supervised learning. To alleviate the problem of data sparsity, there is a growing interest in utilizing the generating capabilities of latent variable models to facilitate data expansion, which is called generative data augmentation. These methods often exploit VAE to learn the hidden semantic representation of each word in a sentence, and use it for subsequent sequence labeling tasks. However, in SLU, in addition to slot filling based on sequence labeling, semantic representation at the utterance level should also be learned and used for intent detection. Furthermore, a great deal of research has been done using VAE for semi-supervised learning. Nevertheless, most of them employ unlabeled data by language modeling. The training of the prediction model still needs a large amount of labeled data, so the sparsity problem cannot be solved fundamentally.

To solve the above challenges, we propose a novel generative model and a semi-supervised learning model for SLU. The motivation of this work comes from the answers to the following questions: (1) How to leverage VAE for joint slot filling and intent detection to alleviate the problem of data sparsity? (2) How to effectively utilize a small amount of labeled data and a large amount of unlabeled data to improve the SLU task?

Specifically, we first propose a novel generative model based on Dirichlet variational autoencoder (DVAE), referred to as DVAE-SLU. Compared with VAE, DVAE exploits a Dirichlet prior for a continuous latent variable that does not suffer from decoder weight collapsing and latent value collapsing [4]. DVAE-SLU uses DVAE to extract more effective utterance semantic representations and generate semantically coherent utterances, thereby improving the performance of SLU with fewer data. Furthermore, based on DVAE-SLU, we propose a semi-supervised learning model for SLU, named SDVAE-SLU. The proposed model can make full use of a small amount of labeled data and a large amount of unlabeled data for training, which greatly reduces the cost of manual labeling. Our main contributions are:

- For the SLU task, we propose a general model DVAE-SLU for generating data augmentation. Unlike previous methods, the proposed method synthesizes complete annotated utterances by using the generative capacity of DVAE. To the best of our knowledge, this is the first work to generate SLU datasets using DVAE.
- To make better use of unlabeled data, we propose a semi-supervised learning model for SLU called SDVAE-SLU. The model can learn both utterance-level and word-level semantic representations for joint slot filling and intent detection.
- The experiments are conducted on the standard ATIS and Snips datasets. Compared with baselines, the enhanced datasets using DVAE-SLU can improve the performance of SLU, especially on small-scale datasets. Additionally, SDVAE-SLU is able to improve SLU by introducing semi-supervised learning.

2 Related Work

Spoken language understanding is one of the hot research fields in NLP. In this section, we introduce several typical studies on slot filling and intent detection, both of which are main sub-tasks of SLU.

For the slot filling task, conventional methods are based on the conditional random field structure, which is good at sequence labeling tasks [17]. In recent years, models based on neural networks have achieved superior performance in several NLP tasks [26-29], gradually replacing traditional models. Yao et al. proposed a method that uses words as input in a recurrent neural network (RNN) language model, and then predicts slot tags instead of words on the output side [18]. Yao et al. proposed a novel model based on long short-term memory (LSTM) for the slot filling task, which has greatly improved compared to RNN [19]. Zhang et al. proposed a novel model that combines the LSTM model with a form of entity location-aware attention, which is more suitable for slot filling [20]. Adel et al. proposed a type-aware convolutional neural network for the slot filling task [21].

For the intent detection task, the conventional approach is to use n-gram as features with universal entities, such as dates and locations [14]. This type of strategy is limited by the dimensionality of the input space. More recently, Hashemi et al. proposed a model based on convolutional neural networks (CNN) to learn query embeddings as features for query intent detection [22]. Balodis et al. proposed an effective intent detection framework that is based on a neural network classifier and word embeddings [23]. Experimental results illustrate that their intent detection model provides state-of-the-art performance. Lin et al. utilized a bidirectional LSTM classifier with margin loss as a feature extractor to

detect unknown intents [24]. By replacing a softmax loss with a margin loss, their method can capture discriminative features by forcing the network to minimize intra-class variance and maximize inter-class variance.

In recent years, neural networks have made a lot of progress in joint learning of slot filling and intent detection. As the utterance of user behavior, slot labels and intent classes can share semantic knowledge with each other. Guo et al. utilized recursive neural networks (RecNN) for joint training of slot filling and intent determination [25]. Liu et al. proposed an attention-based neural network model for joint training [10]. Furthermore, they explored different strategies for integrating alignment information into an encoder-decoder architecture. Li proposed a new joint model, which is based on neural networks and a self-attention mechanism [1]. In their model, the intent-augmented embedding is used as the gate for labeling slot labels. Serdyuk et al. proposed an end-to-end learning system for SLU, which can directly indicate semantic meaning from audio features without textual representations [15]. Xu et al. proposed a new encoder-decoder model with a tag scheme, which unifies two SLU tasks into one sequential labeling task [16].

Although the above methods have made great progress, it is still an open and challenging task for slot filling and intent detection. Therefore, we are motivated to propose an effective method, which can alleviate the problem of data sparsity, thereby improving the performance of SLU systems. To the best of our knowledge, this is the first attempt to utilize DVAE for joint slot filling and intent detection.

3 Methodology

3.1 DVAE-SLU for Data Augmentation

In this paper, vectors are represented by bold letters. A labeled sample of SLU includes an utterance $\boldsymbol{w}$, an equally-long semantic slot sequence $\boldsymbol{s}$ and an intent class y. For a training sample $(\boldsymbol{w}, \boldsymbol{s}, y)$, the objective function is as follows:

$$\mathcal{L}_{SLU}(\psi; \boldsymbol{w}, \boldsymbol{s}, y) = -\log p_{\psi}(\boldsymbol{s}, y \mid \boldsymbol{w}), \tag{1}$$

where ψ represents parameters of the prediction model. Given an utterance $\boldsymbol{w}$, a slot label sequence $\hat{\boldsymbol{s}}$ and an intent class $\hat{y}$ are predicted by maximum the log likelihood: $(\hat{\boldsymbol{s}}, \hat{y}) = \arg\max_{s,y} \log p_{\psi}(\boldsymbol{s}, y \mid \boldsymbol{w})$.

In this subsection, we describe the proposed generative model DVAE-SLU in detail. We start with the standard DVAE and then extend the model by allowing it to generate annotated SLU datasets. DVAE is a deep generative model with a Dirichlet distribution as a prior distribution. Let ϕ be the parameters of the encoder and let θ be the parameters of the decoder. In the SLU data augmentation task, the goal is to maximize the log likelihood of sample $\boldsymbol{w}$ in the corpus $\log p(\boldsymbol{w}) = \log \int p(\boldsymbol{w}, \boldsymbol{z}) d\boldsymbol{z}$, where $\boldsymbol{z}$ is the latent variable of DVAE. However, as the marginalization is difficult to calculate, a proxy network $q_{\phi}(\boldsymbol{z} \mid \boldsymbol{w})$ is introduced in DVAE. Afterwards, based on evidence lower bound (ELBO), we minimize the following objective function:

$$\mathcal{L}_{DVAE}(\theta, \phi; \boldsymbol{w}) = -\mathbb{E}_{z \sim q_{\phi}(z \mid \boldsymbol{w})} \log p_{\theta}(\boldsymbol{w} \mid \boldsymbol{z}) + D_{KL}[q_{\phi}(\boldsymbol{z} \mid \boldsymbol{w}) \,||\, p(\boldsymbol{z})]. \tag{2}$$

Sampling. When optimal parameters θ_O and ϕ_O for $\boldsymbol{w}$ are obtained, we can sample plausible utterances $\hat{\boldsymbol{w}}$ from the variational distribution of $\boldsymbol{w}$ learned by the model:

$$\hat{\boldsymbol{w}} \sim p_{\theta_O, \phi_O}(\boldsymbol{w}) = \int p_{\theta_O}(\boldsymbol{w} \mid \boldsymbol{z}) p_{\theta_O, \phi_O}(\boldsymbol{z}) d\boldsymbol{z}. \tag{3}$$

However, since the true distribution of utterance $\boldsymbol{w}$ is unknown, Equation 3 cannot be solved analytically. Therefore, an approximate method is needed to generate utterances from DVAE. Following [4], we employ a stochastic gradient method by approximating Gamma distributions to infer parameters. DVAE utilizes a Dirichlet distribution conjugated with a multinomial distribution as the prior of latent variables, which is more suitable for the SLU task:

$$\boldsymbol{z} \sim p(\boldsymbol{z}) = Dirichlet(\boldsymbol{\alpha}), \boldsymbol{w} \sim p_\theta(\boldsymbol{w} \mid \boldsymbol{z}), \tag{4}$$

where $\boldsymbol{\alpha}$ denotes Dirichlet hyperparameters. The approximate posterior distribution $q_\phi(\boldsymbol{z} \mid \boldsymbol{w})$ in the encoder is sampled from $Dirichlet(\hat{\boldsymbol{\alpha}})$. The approximate posterior parameter $\hat{\boldsymbol{\alpha}}$ is obtained by the multilayer perceptron of utterances $\boldsymbol{w}$ with a softplus output function. Instead of sampling $\boldsymbol{z}$ directly from a Dirichlet distribution, DVAE takes advantage of the fact that the Dirichlet distribution can be composed of multiple Gamma random variables to sample $\boldsymbol{z}$ using a Gamma composition method. Specifically, we first draw $\boldsymbol{v} \sim MultiGamma(\boldsymbol{a}, \beta, \boldsymbol{1}_K)$ where $MultiGamma(\boldsymbol{a}, \beta, \boldsymbol{1}_K)$ denotes K independent random variables that follow Gamma distributions. Secondly, $\boldsymbol{v}$ is normalized by its summation $\sum v_i$ and the objective function is as follow:

$$\mathcal{L}(\theta, \phi; \boldsymbol{w}) = \mathbb{E}_{z \sim q_\phi(z|w)} \log p_\theta(\boldsymbol{w}|\boldsymbol{z}) - (\sum \log \Gamma(a_k) - \sum \log \Gamma(\widehat{a_k}) + \sum \log \Gamma(\widehat{a_k} - a_k) \Psi(\widehat{a_k})), \tag{5}$$

where $\Psi(\bullet)$ is a digamma function. The approximate method requires the configuration of the stochastic gradient variational Bayes estimator on the Dirichlet distribution. Since the Dirichlet distribution is a composition of Gamma random variables, our model utilizes asymptotic approximation to approximate the inverse Gamma cumulative distribution function (CDF). Specifically, if $F(v; \alpha, \beta)$ denote a CDF of random variable $\boldsymbol{v}$ and $\boldsymbol{v} \sim Gamma(\alpha, \beta)$, we can approximate the inverse CDF as $F^{-1}(u; \alpha, \beta) \approx \beta^{-1}(ua\Gamma(\alpha))^{1/\alpha}$ [5]. Therefore, an auxiliary variable $u \sim Uniform(0,1)$ is introduced to replace all the randomness of $\boldsymbol{v}$, and Gamma random variable $\boldsymbol{v}$ can be seen as a deterministic value about α and β. The sampling process of DVAE-SLU is as follows:

- Given the number of sampled utterances n, initialize an empty list $\mathbb{M}$;
- When the number of sampled utterances is less than n, the following steps are repeated:
  sample a real utterance $\boldsymbol{w}$;
  estimate $\boldsymbol{z}$ by approximation with inverse Gamma CDF;
  sample $\hat{\boldsymbol{w}}$ from the likelihood $p_\theta(\boldsymbol{w} \mid \boldsymbol{z})$;
  append $\hat{\boldsymbol{w}}$ to list $\mathbb{M}$.

Generative data augmentation process. After obtaining generated utterances, DAVE-SLU extends the DVAE model by predicting slot tags and the intent of an utterance. The generation of slot tags and intent classes depends on the hidden variable $\boldsymbol{z}$ and generated utterances $\hat{\boldsymbol{w}}$. The modified objective function for the SLU task is as follows:

$$\mathcal{L}(\phi, \psi; \boldsymbol{w}, \boldsymbol{s}, y) = \mathbb{E}_{\mathbf{z} \sim q_\phi}[\log p_\psi(\boldsymbol{s}, y \mid \hat{\boldsymbol{w}}, \boldsymbol{z})], \tag{6}$$

where ϕ denotes the optimal parameter of data augmentation process on the original dataset, and ψ denotes the optimal parameter of slot filling and intention detection on the generated dataset $\hat{\boldsymbol{w}}$. Considering Dirichlet variational autoencoders for data augmentation (Equation 2) and spoken language understanding (Equation 6), the joint training objective function of DVAE -SLU is as follows:

$$\begin{aligned} \mathcal{L}_{DVAE+SLU}(\theta, \phi, \psi; \boldsymbol{w}, \boldsymbol{s}, y) = & \mathbb{E}_{\mathbf{z} \sim q_\phi}[\log p_\theta(\boldsymbol{w} \mid \boldsymbol{z})] \\ & + \mathbb{E}_{\mathbf{z} \sim q_\phi}[\log p_\psi(\boldsymbol{s}, y \mid \hat{\boldsymbol{w}}, \boldsymbol{z})] \\ & - D_{KL}[q_\phi(\boldsymbol{z} \mid \boldsymbol{w}) \| p_\theta(\boldsymbol{z} \mid \boldsymbol{w})] \end{aligned} . \tag{7}$$

The DVAE-SLU model can be divided into two modules. One is a data augmentation module that samples hidden variables and generates plausible utterances by DVAE. The other is a prediction module that employs generated utterances to improve the SLU task. From the perspective of an encoder-decoder framework, the data augmentation module belongs to the encoder, while the prediction module belongs to the decoder. The structure of the proposed model is shown in Fig. 1.

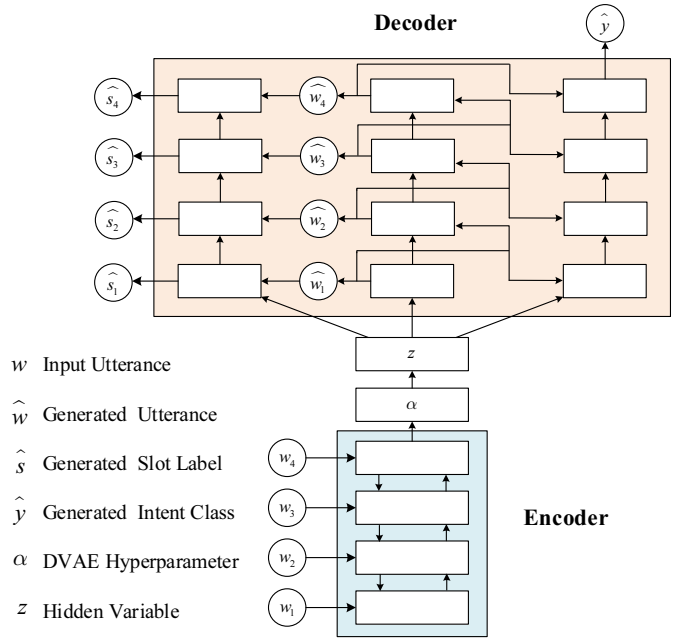

Fig. 1. The structure of DVAE-SLU

3.2 SDVAE-SLU for Semi-supervised Learning

In this subsection, a series of structural modifications are made to the DVAE-SLU model, and we propose a novel SLU model SDVAE-SLU that can be used for semi-supervised joint semantic extraction. Kingma et al. first proposed a semi-supervised learning framework based on variational inference in 2014 [3]. This method proposed objective functions for both labeled and unlabeled data. Taking the intent detection task as an example, for labeled data, given the pair $(\boldsymbol{w}, y)$ composed of utterance $\boldsymbol{w}$ and intent y, the relationship between ELBO and the corresponding hidden variable $\boldsymbol{\eta}$ is:

$$\begin{aligned}\log p_\theta(\boldsymbol{w}, y) \geq \mathbb{E}_{\eta\sim q_\phi(\boldsymbol{\eta}|\boldsymbol{w},y)}\left[\log p_\theta(\boldsymbol{w}\mid y,\boldsymbol{\eta})\right] - D_{KL}(q_\phi(\boldsymbol{\eta}\mid\boldsymbol{w},y)\,\|\,p(\boldsymbol{\eta}))\\ + \log p_\theta(y) = -\mathcal{L}(\boldsymbol{w}, y)\end{aligned}, \tag{9}$$

where the first term is the expectation of the log-likelihood of the hidden variable $\boldsymbol{\eta}$, the second term is the Kullback-Leibler (KL) distance between prior distribution $p(\boldsymbol{\eta})$ and posterior distribution $q_\phi(\boldsymbol{\eta}\mid\boldsymbol{w}, y)$.

For unlabeled data, their intent is predicted by a classifier $q_\phi(y\mid\boldsymbol{w})$, and its variation lower bound is:

$$\log p_\theta(\boldsymbol{w}) \geq \sum_y q_\phi(y\mid\boldsymbol{w})\left(-\mathcal{L}(\boldsymbol{w}, y)\right) + \mathcal{I}\left(q_\phi(y\mid\boldsymbol{w})\right) = -\mathcal{K}(\boldsymbol{w}), \tag{10}$$

where $\mathcal{I}\left(q_\phi(y\mid\boldsymbol{w})\right)$ represents a hypothesized distribution learned by the classifier from data. The objective function of the entire dataset is:

$$J = \sum_{(\boldsymbol{w},y)\in\mathcal{D}_l}\mathcal{L}(\boldsymbol{w}, y) + \sum_{\boldsymbol{w}\in\mathcal{D}_u}\mathcal{K}(\boldsymbol{w}) + \lambda\mathbb{E}_{(\boldsymbol{w},y)\in\mathcal{D}_l}\left[-\log q_\phi(y\mid\boldsymbol{w})\right], \tag{11}$$

where D_l denotes the labeled dataset, D_u denotes the unlabeled dataset, λ is a trade-off hyperparameter, and the third term represents the loss of the classifier.

The structure of the semi-supervised variational autoencoder for intent detection is shown in Fig. 2. The model contains three main parts: the encoder network $q_\phi(\boldsymbol{\eta}\mid\boldsymbol{w}, y)$, the decoder network $p_\theta(\boldsymbol{w}\mid y,\boldsymbol{\eta})$, and the semi-supervised classifier $q_\phi(y\mid\boldsymbol{w})$.

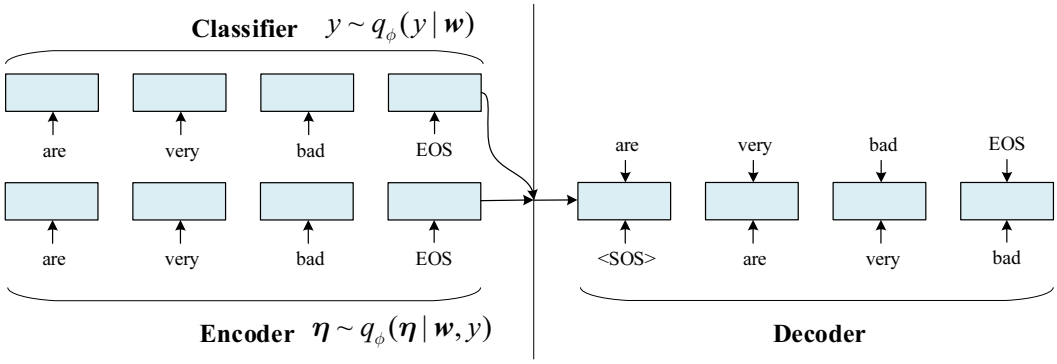

Fig. 2. The structure of the semi-supervised variational autoencoder

DVAE-SLU directly reconstructs the entire input sequence $\boldsymbol{w}$ by hidden variable $\boldsymbol{z}$ and generates new utterances $\hat{\boldsymbol{w}}$. However, since the reconstruction process focuses on the utterances level, the improvement of slot filling is not as significant as intent detection. Therefore, SDVAE-SLU extends the DVAE-SLU model by introducing semi-supervised sequence labeling [6] and modifying the reconstruction process, which can effectively improve the performance of slot filling. The structure of SDVAE-SLU is shown in Fig. 3.

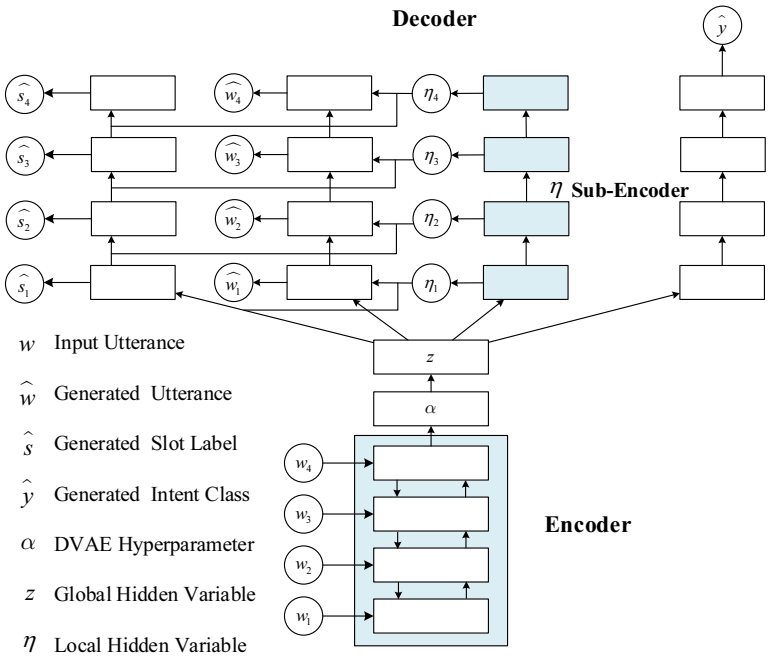

Fig. 3. The structure of SDVAE-SLU

As shown in Fig. 3, the encoder and $\boldsymbol{z}$ sampling process of SDVAE-SLU are consistent with DVAE-SLU. In the decoder, SDVAE-SLU adds a sub-encoder with a similar structure to sample the hidden variable $\boldsymbol{\eta}$. After the encoder samples the utterance-level global hidden variable $\boldsymbol{z}$, it concatenates each word w_t in the utterance with the corresponding label s_t to form a new vector. The vector is used as the input of the sub-encoder and sampled to obtain the local hidden variable η_t for each word w_t. The generation network of the decoder uses η_t to generate the reconstructed $\hat{w}_t$. According to the above generation process, the variational lower bound can be derived as follows:

$$\begin{aligned}\log p\left(\widehat{w}_t \mid w_t\right) &\geq \mathbb{E}_{\eta_t, z\sim q_\phi(\eta_t,z)} \log p_\theta\left(\widehat{w}_t \mid \eta_t, \boldsymbol{z}\right) - \log\frac{q_\phi\left(\eta_t \mid \boldsymbol{w}\right)}{p_\theta\left(\eta_t\right)} - \log\frac{q_\phi\left(\boldsymbol{z} \mid \boldsymbol{w}\right)}{p_\theta\left(\boldsymbol{z}\right)} \\ &= \mathbb{E}_{\eta_t, z\sim q_\phi(\eta_t,z)} \log p_\theta\left(\widehat{w}_t \mid \eta_t, \boldsymbol{z}\right) - D_{KL}[q_\phi\left(\eta_t \mid \boldsymbol{w}\right) \| \, p_\theta\left(\eta_t\right)] \\ &\quad - D_{KL}[q_\phi\left(\boldsymbol{z} \mid \boldsymbol{w}\right) \| \, p_\theta\left(\boldsymbol{z}\right)]\end{aligned} . \tag{12}$$

The first term in Equation 12 represents the loss of the reconstruction process from $w_t \sim z \sim \eta_t \sim \widehat{w}_t$. The second and third terms respectively represent the KL distance between the approximate distribution and the real distribution during two sampling processes. By extending Equation 12 to the whole sequence, the objective function of the whole utterance $\boldsymbol{w}$ in the generation model can be obtained as follows:

$$\begin{aligned}\mathcal{L}_{gm} = \mathcal{L}_{gm}(\theta, \phi; \boldsymbol{w}) = \sum_{t=1}^{T} \mathbb{E}_{\eta_t, z\sim q_\phi(\eta_t,z)} \log p_\theta\left(\widehat{w}_t \mid \eta_t, \boldsymbol{z}\right) \\ - D_{KL}[q_\phi\left(\boldsymbol{\eta} \mid \boldsymbol{w}\right) \| \, p_\theta\left(\boldsymbol{\eta}\right)] - D_{KL}[q_\phi\left(\boldsymbol{z} \mid \boldsymbol{w}\right) \| \, p_\theta\left(\boldsymbol{z}\right)]\end{aligned} . \tag{13}$$

Based on the semi-supervised learning structure in [6], for the slot filling task, the objective function (for each word w_t) with labeled data $\mathcal{D}_l$ is as follows:

$$\mathcal{L}_{(w_t, s_t)\in\mathcal{D}_l}\left(w_t, s_t\right) = \mathbb{E}_{\eta_t\sim q_\phi(\eta_t|z)}\left[-\log q_\phi(s_t \mid \eta_t)\right]. \tag{14}$$

For a large amount of unlabeled data, the prediction labels of the classifier are used to guide learning. The cross-entropy of the predicted distribution and real samples is calculated to increase the prediction confidence of the classifier corresponding to the labeled data. That is:

$$\mathcal{L}_{sf} = \mathcal{L}_{(w_t, s_t)\in\mathcal{D}_l}\left(w_t, s_t\right) + \lambda_1 \mathbb{E}_{\eta_t\sim q_\phi(\eta_t|z)}\left[\sum_{s_t} q_\phi(s_t \mid \eta_t) \log q_\phi(s_t \mid \eta_t)\right], \tag{15}$$

where λ_1 denotes a trade-off hyperparameter. During semi-supervised training, λ_1 is tuned based on a validation set. The first term in Equation 15 is a labeled data loss function, and the second term is an unlabeled data loss function. The semi-supervised classifier loss function for the intent detection task is similar to the slot filling task, as follows:

$$\mathcal{L}_{id} = \mathbb{E}_{z\sim q_\phi(z|w)}\left[-\log q_\phi\left(y \mid \boldsymbol{z}\right)\right] + \lambda_2 \mathbb{E}_{z\sim q_\phi(z|w)}\left[\sum_{y} q_\phi\left(y \mid \boldsymbol{z}\right) \log q_\phi\left(y \mid \boldsymbol{z}\right)\right]. \tag{16}$$

Similarly, in Equation 16, the first term is a labeled data loss function, and the second term is an unlabeled data loss function. Synthesized Equation 13, 15 and 16, the optimization objective function of the SDVAE-SLU model for each input sequence $\boldsymbol{w}$ is as follows:

$$\mathcal{L}_{SDVAE-SLU} = \mathcal{L}_{gm} + \sum_{t=1}^{T} \mathcal{L}_{sf} + \mathcal{L}_{id}. \tag{17}$$

4 Experiments

In this section, we carry out extensive experiments on two real-world SLU datasets to evaluate DVAE-SLU and SDVAE-SLU. The experimental results show that DVAE-SLU provides a promising improvement for existing SLU models, and the effectiveness of the SDVAE-SLU model.

4.1 Dataset and Setting

We evaluate our models on the following two datasets:

- **ATIS:** In the SLU task, Airline Travel Information System (ATIS) is a classic dataset [7]. It can provide an ideal comparison environment for experiments. The dataset contains text data converted

from recordings when users book flights. The training set consists of 4,478 utterances and the testing set consists of 500 utterances. Furthermore, ATIS contains 120 semantic slot labels and 21 intents.

- **Snips:** The Snips dataset is an open source natural language comprehension evaluation dataset. It contains various intentional queries such as manipulating playlists and booking restaurants collected by virtual assistants. The size of each intention in the Snips is roughly the same. The training set contains 13,084 utterances and the testing set contains 700 utterances. The dataset contains 72 semantic slot labels and 7 intents.

Tables 1 and 2 show detailed descriptions of the two datasets.

Table 1. Statistics of two datasets

	Snips	ATIS
#Train	13,084	4,478
#Val	700	500
#Test	700	893
#Slot	72	120
#Intent	7	21
Vocabulary	11,241	722

Table 2. Examples of two datasets

	ATIS
Utterance	1. what flights are available from pittsburgh to baltimore on thursday morning
	2. cheapest airfare from tacoma to orlando
Slot	1. O O O O O B-fromloc.city_name O B-toloc.city_name O B-depart_date.day_name B-depart_time.period_of_day
	2. O B-cost_relative O O B-fromloc.city_name O B-toloc.city_name O
Intent	1. flight
	2. airfare
	Snips
Utterance	1. What is the forecast for 8/26/2022 in Vermont
	2. Play songs on Itunes
Slot	1. O O O O O B-timeRange O B-state
	2. O O O B-service
Intent	1. GetWeather
	2. PlayMusic

In comparison, the Snips dataset has a larger vocabulary size, more user intent categories, and almost the same number of each intent. The vocabulary involved in the ATIS dataset is related to aeronautical information and is much smaller than Snips. In addition, about 74% of the statements in the ATIS dataset are intended for flight.

In the DVAE-SLU model, we employ a pre-trained FastText[1] model to initialize word embeddings, which has a 300-dimensional embedding containing 2 million words trained on Wikipedia [8]. The embedding dimensions of slot labels and intents are set to 200 and 100, respectively. The word, slot label and intent embeddings are trained jointly with the network. For DVAE, we set $\boldsymbol{a} = 0.99 \cdot \boldsymbol{I}_{100}$ and $\beta = 1$.

For the generative model, the encoder network leverages a single-layer bidirectional LSTM [9] to encode word embeddings in both directions, and generates final hidden states $(h_1^{encoder}, ..., h_T^{encoder})$ by a max-pooling layer, whose dimension is set to 256. The three networks of the decoder all use a single-layer unidirectional LSTM. The dimension of the hidden outputs of a word, a slot label and an intent are set to 1024. In the decoding process, a beam search algorithm is used to find the most likely candidate sequences from the conditional distribution, and the window size of beam search is set to 15.

In the training process, in order to effectively train the model, we utilize a teacher-forcing strategy to train the SLU network on the ground truth $\boldsymbol{w}$ instead of the prediction sequence $\widehat{\boldsymbol{w}}$. In addition, to prevent overfitting, the dropout rate is set to 0.5. Adam optimizer is used for optimization, and the initial

[1] https://fasttext.cc/docs/en/english-vectors.html

learning rate is set to 0.001.

The SDVAE-SLU model is an extension based on DVAE-SLU. For word embeddings, the FastText model is also used to initialize word vectors. The dimensions of the word, slot label and intent embeddings are unchanged. The sub-encoder for η sampling has the same structure as the global encoder for z, which is implemented using an LSTM network with a max pooling layer.

4.2 Evaluation of DVAE-SLU

The base models used in the experiments include the attention-based encoder-decoder model (AttED) [10] and the Slot-Gated model [11]. With open source code, we can reproduce experimental results similar to the base models. The difference in results caused by different data preprocessing methods will not affect this experiment. The performance of the SLU model can be evaluated by the following metrics: (1) F1-score of slot filling; (2) F1-score of intent detection; (3) comprehensive semantic accuracy. The comprehensive semantic accuracy is obtained by adding the correct number of semantic slots to the correct number of intention detection, and then dividing by (sequence size +1).

The first experiment is to simulate data augmentation performance in a data scarce scenario. For comparison, we utilize a labeled data generation model DeepLSTM as a baseline method [12]. DVAE-SLU and DeepLSTM learn from the encoded information of input utterances, and then decode reconstructed utterances and the corresponding semantic slot labels. To simulate the environment where data is scarce, we divide the ATIS dataset into data subsets of three sizes: full / medium / small. The small-scale dataset randomly divides the training set into 40 parts of the same size, each with about 111 ~ 112 utterances. The medium-sized dataset randomly divides the training set into 10 parts of the same size, each with about 447 ~ 448 utterances. The full-scale dataset is a complete training set with 4,478 utterances. For the full-scale dataset, the proposed model generates 11,000 generated utterances and adds them to the enhanced dataset $\mathbb{M}$. For medium and small datasets, the utterances are generated by repeated running the model for 30 times or 110 times, respectively.

Table 3 shows the performance of SLU after data augmentation using three different scale datasets. In this experiment, AttED is used as a base model. Since DeepLSTM only generates slot labels, it does not experiment on intent detection and comprehensive semantic accuracy. From the results, it can be seen that after using DVAE-SLU for data augmentation, the SLU performance is better than the base model. The slot filling performance of our model is also slightly better than using the DeepLSTM model. This advantage is more significant on small and medium-scale datasets. The reason is that DVAE employs a Dirichlet prior for latent variables, which in turn generate more semantically coherent utterances. The results validate the effectiveness of DVAE-SLU against baselines in generating SLU datasets. For the full-scale dataset, the performance improvement achieved by the proposed model is relatively small, which may be caused by the high homogeneity of the ATIS dataset. The dataset with similar features makes the sampling space smaller, resulting in little improvement.

Table 3. Data augmentation results for the ATIS dataset

Model	Slot Filling (F1)			Intent (F1)			Semantic (Acc.)		
	Small	Med.	Full	Small	Med.	Full	Small	Med.	Full
AttED	72.46	88.17	95.34	82.57	90.52	**97.22**	35.02	65.14	85.93
AttED+DeepLSTM	74.80	89.10	95.24						
AttED+DVAE-SLU	**74.87**	**89.23**	**95.38**	**83.10**	**90.81**	97.20	**36.61**	**66.64**	**85.97**

The second experiment is the performance of data augmentation under different datasets. Datasets for comparison include ATIS and Snips, and the base model is Slot-Gated [11]. Complete training sets are used for training in this experiment.

Table 4 illustrates the performance of DVAE-SLU on different datasets by using the Slot-Gated SLU models. Datasets are augmented (prefixed by +) using DVAE-SLU. Experimental results show that both datasets achieved improvement in SLU after using the proposed model for data augmentation. The Snips dataset with richer vocabulary and more differences between intent statements achieved better slot filling performance improvement than the ATIS dataset. This finding confirms the positive correlation between the complexity of the dataset and the improvement in the SLU task.

Table 4. The performance of DVAE-SLU on different datasets by using the Slot-Gated SLU model

Dataset	Slot Filling (F1)	Intent (F1)	Semantic (Acc.)
ATIS	94.9	95.0	84.2
ATIS+	**95.9**	**95.9**	**85.7**
Snips	88.2	97.0	74.6
Snips+	**89.0**	**97.6**	**76.3**

4.3 Evaluation of SDVAE-SLU

Due to the small number of the ATIS dataset, it is difficult to simulate the actual scene of a small amount of labeled data and a large amount of unlabeled data. Therefore, we combine the MIT Restaurant (MR) dataset, the MIT Movies dataset (MM) [13] and the ATIS dataset into a larger dataset. The training set of the combined dataset contains 30,299 (25,821 new) queries, 191 (71 new) semantic slot labels, and the vocabulary size is 16,049 (15,327 new). The testing set contains 6,810 queries. In addition, considering that intents in MR and the MM are close to each other, they correspond to one intention and are added to the combined dataset, with a total of 23 intention labels.

For semi-supervised learning, we randomly selected 5,000, 10,000, 15,000, 20,000 and 25,000 statements from the combined dataset as labeled datasets, and the remaining statements as unlabeled datasets. For each labeled dataset, we randomly choose 80% as the training set and the rest as the validation set. For example, in the 10K dataset, 8,000 labeled statements are used for training, 2,000 labeled statements are used for verification, and the remaining 20,299 statements are treated as unlabeled data.

The first experiment is to verify the SLU performance of the SDVAE-SLU model on the combined datasets with different annotation scales.

Table 5 shows the performance of SDVAE-SLU with different annotation scales. In Table 5, “Label” means using labeled data of a specified size for supervised training, while SDVAE-SLU introduces remaining unlabeled data for semi-supervised training. It can be seen from the experimental results that after introducing semi-supervised learning, the SLU performance has been significantly improved under different scale datasets. This observation suggests that the SVSAE-JLU model can learn the distribution of the data from the unlabeled data, and then use the prediction results to guide the task of slot filling and intent detection.

Table 5. The performance of SDVAE-SLU with different annotation scales

Data scale	Slot Filling (F1)		Intent (F1)	
	Label	SDVAE-SLU	Label	SDVAE-SLU
5K	67.35	**68.49**	89.51	**92.58**
10K	71.07	**71.78**	90.13	**92.95**
15K	73.28	**73.93**	91.36	**94.13**
20K	74.59	**75.20**	92.47	**94.33**
25K	75.92	**76.34**	93.68	**94.69**
ALL	76.70	-	94.74	-

The second experiment compares the performance of slot filling between semi-supervised joint semantic extraction SDVAE-SLU and semi-supervised sequence labeling. Specifically, the DVAE is replaced with a semi-supervised sequence labeling RNN model and implemented using LSTM, denoted as SRNN-SLU. In the SRNN-SLU model, the intent detection and slot filling tasks are independent of each other. Experiments are performed using the ATIS dataset. When the labeled data accounted for 10%, 20%, 30%, 50% and 70%, the semi-supervised SLU performance of the proposed model and the SRNN-SLU model are compared.

Table 6 illustrates the semi-supervised SLU performance of SDVAE-SLU and SRNN-SLU with different annotation scales. From the experimental results, it can be seen that compared with the general semi-supervised sequence labeling model RNN, SDVAE-SLU can explicitly associate the intent type with the semantic slot type through the attention mechanism and the gate structure. Therefore, in the case of sufficient labeled data, the proposed model is better than the simple semi-supervised sequence labeling

model.

Table 6. The semi-supervised SLU performance of SDVAE-SLU and SRNN-SLU

Labeled scale	Slot Filling (F1)	
	SRNN-SLU	SDVAE-SLU
10%	89.07	**89.21**
20%	90.57	**91.17**
30%	91.68	**92.14**
50%	92.94	**93.82**
70%	94.13	**94.91**
ALL	94.95	**95.37**

5 Conclusion

In this paper, based on Dirichlet variational autoencoder, we first propose a new generative data augmentation model DVAE-SLU for spoken language understanding. DVAE-SLU is able to generate both spoken utterances and their semantic labels (slot labels and intent classes). Furthermore, based on the DVAE-SLU model, a semi-supervised learning model for joint slot filling and intent detection is proposed, called SDVAE-SLU. The proposed model can naturally combine labeled and unlabeled data, thereby improving the performance of the SLU task. Experiments on two real-world datasets validate the effectiveness of our proposed models. In the future, we will improve our models by introducing additional knowledge.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This work was supported in part by Doctoral Start-up Fund of Jianghan University (No. 1028/06060001) and Application Foundation Frontier Project of Wuhan (NO. 2020010601012289). We would like to thank Yan Xiao and Ping Dong for their efforts in the data processing.

References

[1] C. Li, L. Li, J. Qi, A self-attentive model with gate mechanism for spoken language understanding, in: Proc. 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[2] Y. Shin, K. Yoo, S. Lee, Utterance generation with variational auto-encoder for slot filling in spoken language understanding, IEEE Signal Processing Letters 26(3)(2019) 505-509.

[3] D. Kingma, M. Welling, Auto-encoding variational bayes, in: Proc. 2014 International Conference on Learning Representations, 2014.

[4] W. Joo, W. Lee, S. Park, I. Moon, Dirichlet variational autoencoder, Pattern Recognition 107(49)(2020) 514-524.

[5] D. Knowles, Stochastic gradient variational bayes for gamma approximating distributions, <http://arxiv.org/abs/1509.01631v1>, 2015.

[6] M. Chen, Q. Tang, K. Livescu, K. Gimpel, Variational sequential labelers for semi-supervised learning, in: Proc. 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

[7] C. Hemphill, J. Godfrey, G. Doddington, The ATIS spoken language systems pilot corpus, in: Proc. 1990 Speech and Natural Language Workshop, 1990.

[8] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proc. 2018 International Conference on Language Resources and Evaluation, 2018.

[9] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, A. Bordes, Supervised learning of universal sentence representations from natural language inference data, in: Proc. 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

[10] B. Liu, I. Lane, Attention-based recurrent neural network models for joint intent detection and slot filling, in: Proc. 2016 Conference on Interspeech, 2016.

[11] W. Fedus, I. Goodfellow, A. Dai, Maskgan: Better text generation via filling in the _, in: Proc. 2018 International Conference on Learning Representations, 2018.

[12] G. Kurata, B. Xiang, B. Zhou, Labeled data generation with encoder-decoder LSTM for semantic slot filling, in: Proc. 2016 Conference on Interspeech, 2016.

[13] J. Liu, P. Pasupat, S. Cyphers, J. Glass, Asgard: A portable architecture for multilingual dialogue systems, in: Proc. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[14] X. Zhang, H. Wang, A joint model of intent determination and slot filling for spoken language understanding, in: Proc. 2016 International Joint Conference on Artificial Intelligence, 2016.

[15] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, Y. Bengio, Towards end-to-end spoken language understanding, in: Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, 2018.

[16] C. Xu, Q. Li, D. Zhang, J. Cui, Z. Sun, H. Zhou, A model with length-variable attention for spoken language understanding, Neurocomputing 379(28)(2020) 197-202.

[17] C. Raymond, G. Riccardi, Generative and discriminative algorithms for spoken language understanding, in Proc. 2017 Annual Conference of the International Speech Communication Association, 2007.

[18] K. Yao, G. Zweig, M. Hwang, Y. Shi, D. Yu, Recurrent neural networks for language understanding, in: Proc. 2013 Conference on Interspeech, 2013.

[19] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, Y. ShiK, Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, Y. Shi, Spoken language understanding using long short-term memory neural networks, in: Proc. 2014 IEEE Spoken Language Technology Workshop, 2014.

[20] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. Manning, Position-aware attention and supervised data improve slot filling, in: Proc. 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

[21] H. Adel, H. Schütze, Type-aware convolutional neural networks for slot filling, Journal of Artificial Intelligence Research 66(11)(2019) 297-339.

[22] H. Hashemi, A. Asiaee, R. Kraft, Query intent detection using convolutional neural networks, in: Proc. 2016 International Conference on Web Search and Data Mining, Workshop on Query Understanding, 2016.

[23] K. Balodis, D. Deksne, Intent detection system based on word embeddings, in: Proc. 2018 International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2018.

[24] T. Lin, H. Xu, Deep unknown intent detection with margin loss, in Proc. 2019 Annual Meeting of the Association for Computational Linguistics, 2019.

[25] D. Guo, G. Tur, W. Yih, G. Zweig, Joint semantic utterance classification and slot filling with recursive neural networks, in: Proc. 2014 IEEE Spoken Language Technology Workshop, 2014.

[26] W. Gao, M. Peng, H. Wang, Y. Zhang, Q. Xie, G. Tian, Incorporating word embeddings into topic modeling of short text, Knowledge and Information Systems 61(2)(2019) 1123-1145.

[27] B. Hu, Y. Liu, M. Sun, K. Mi, An extended attention-based LSTM with knowledge embedding for aspect-level sentiment analysis, Journal of Computers 30(4)(2019) 176-184.

[28] W, Gao, M. Peng, H. Wang, Y. Zhang, W. Han, G. Hu, Q. Xie, Generation of topic evolution graphs from short text streams, Neurocomputing 383(28)(2020) 282-294.

[29] F. Yin, H. Xu, H. Gao, M, Bian, Research on weibo public opinion prediction using improved genetic algorithm based BP neural networks, Journal of Computers 30(3)(2019) 82-101.