A Structured Latent Variable Recurrent Network With Stochastic Attention For Generating Weibo Comments

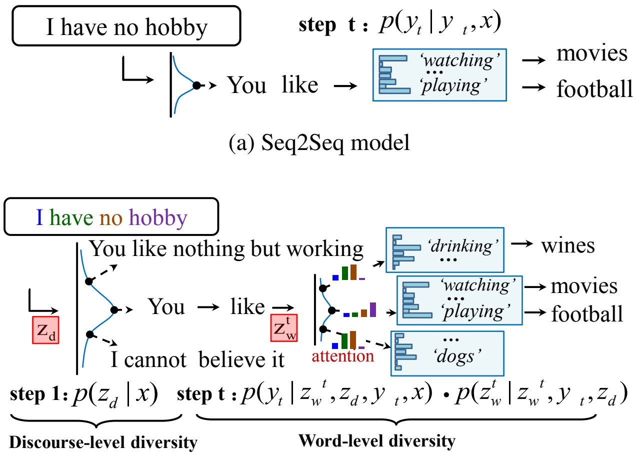
Shijie Yang[1] , **Liang Li**[2*] , **Shuhui Wang**[2] , **Weigang Zhang**[3] , **Qingming Huang**[1,2,4] and **Qi Tian**[5]
[1]University of Chinese Academy of Sciences, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences, China
[3]School of Computer Science and Technology, Harbin Institute of Technology, China
[4]Peng Cheng Laboratory, China
[5]Noah's Ark Lab, Huawei Technologies, China
shijie.yang@vipl.ict.ac.cn, {liang.li, wangshuhui}@ict.ac.cn, wgzhang@hit.edu.cn,
qmhuang@ucas.ac.cn, tian.qi1@huawei.com

Abstract

Building intelligent agents to generate realistic Weibo comments is challenging. For such realistic Weibo comments, the key criterion is improving diversity while maintaining coherency. Considering that the variability of linguistic comments arises from multi-level sources, including both discourse-level properties and word-level selections, we improve the comment diversity by leveraging such inherent hierarchy. In this paper, we propose a structured latent variable recurrent network, which exploits the hierarchical-structured latent variables with stochastic attention to model the variations of comments. First, we endow both discourse-level and word-level latent variables with hierarchical and temporal dependencies for constructing multi-level hierarchy. Second, we introduce a stochastic attention to infer the key-words of interest in the input post. As a result, diverse comments can be generated with both discourse-level properties and local-word selections. Experiments on open-domain Weibo data show that our model generates more diverse and realistic comments.

1 Introduction

Generating realistic comments for social applications is attaching lots of attention [Shen *et al.*, 2019; Holtzman *et al.*, 2019; Zhang *et al.*, 2018]. Compared with rule-based [Goddeau *et al.*, 1996] and retrieval-based methods [Eric and Manning, 2017], generative sequence-to-sequence (Seq2Seq) [Sutskever *et al.*, 2014; Liu *et al.*, 2018; Yang *et al.*, 2019; Zha *et al.*, 2019] models using encoder-decoder architectures have been widely used due to their super capacity for sequential data. As Figure 1a, traditional Seq2Seq models directly maximize the likelihood between input and output. They tend to generate "safe" and meaningless comments of high-frequency, and such comments lack the diversity [Li *et al.*, 2016]. Closely related to response generation, generating diverse and coherent comments is important for improving the user experience of intelligent agents.



(a) Seq2Seq model

(b) Our proposed SARN

Figure 1: Illustrations of comment generation process. (a) Traditional Seq2Seq models predict the t-th word y_t via $p(y_t|\mathbf{y}_{<t}, \mathbf{x})$ given input $\mathbf{x}$, where stochastic variation is parameterized by this conditional output distribution. (b) We introduce latent variables $\mathbf{z}_d$ and $\mathbf{z}_w$ with *stochastic attention* to capture the discourse-level and word-level stochastic variations. Our model estimates the distribution of $p(y_t|\mathbf{z}_w^{\leq t}, \mathbf{z}_d, \mathbf{y}_{<t}, \mathbf{x})$. Diverse comments are produced conditioned on different assignments of latent variables.

Recently, some methods are proposed to improve comment diversity, which can be divided into two groups. The first one focuses on improving the word generation process, including designing new variations of generation procedures and finding new types of decoding objectives. [Mou *et al.*, 2016] proposed a backward-forward decoding process which generated a reply start from diversified middle keyword. [Vijayakumar *et al.*, 2016] and [Shao *et al.*, 2017] proposed diversity-enforced beam-search procedures. Moreover, [Yao *et al.*, 2016] incorporated Inverse Document Frequency into the decoder objective to promote the comment diversity. The

[*]Corresponding Author

second group emphasizes that comments are diversified in terms of high-level properties or attributes, such as language style, topic, intention. These properties are explored as context information to generate more specific comments. [Xing *et al.*, 2017] fed topic information into a joint attention module to generate topic-related comments. [Zhou *et al.*, 2017b] added emotion context to build emotional chatting agents. Moreover, another series of works implicitly modeled the high-level attributes using stochastic latent variables [Zhao *et al.*, 2017; Cao and Clark, 2017; Serban *et al.*, 2017; Shen *et al.*, 2017; Liu *et al.*, 2019]. In such way, diverse comments are able to be sampled conditioned on the latent variables. However, most of above models treat comment generation as a single-level process, neglecting the hierarchy in natural comments.

In practice, the variabilities of comments arise from such the hierarchy, including both global discourse-level properties and word-level selections. Given one post, first, a number of comments can be generated conditioned on **discourse-level diversity** (i.e. different emotions or language styles). Then, for one specific discourse-level selection, various local word-trajectories are able to be generated conditioned on **word-level diversity**. How to formulate the above procedure remains an open question.

In this paper, we propose a *structured latent variable Recurrent Network with Stochastic Attention* (SARN), a probabilistic model that exploits both hierarchical-structured latent variables and the stochastic attention to promote multi-level diversity of comments. First, we introduce both discourse-level variable and word-level latent variables to build the hierarchy of comments. In detail, the discourse-level variable is used to capture the high-level properties such as style, topic and intention. Word-level variables are leveraged by the stochastic attention to infer the variations of the key-words of interest in the input post. Second, we endow these latent variables with hierarchical and temporal dependencies, which are specifically designed to model the complex structure of comments. As result, diverse comments are able to be produced conditioned on different assignments of these latent variables. The proposed SARN is trained using the Stochastic Gradient Variational Bayesian framework [Kingma and Welling, 2013] which maximizes the variational lower bound of the conditional log likelihood. Experiments on open-domain Weibo dataset showed that the proposed SARN yields significantly more diverse comments at both discourse-level and word-level compared to other methods.

2 Preliminary

Given an input $\mathbf{x} = (x_1, ..., x_T)$ and a comment sequence $\mathbf{y} = (y_1, ..., y_{T'})$, where $y_t$is the t-th caption word, Seq2Seq model is trained to maximize the probability of $p(\mathbf{y}|\mathbf{x})$. A bidirectional recurrent network (B-RNN) first summarizes the input sequence $\mathbf{x}$ into hidden states $\mathbf{h}_t$.

Soft-attention. A decoder RNN calculates the context $\mathbf{c}_t$, and estimates the probability of $\mathbf{y}$. In detail, the decoder RNN takes a context $\mathbf{c}_t$ and the previously decoded word $y_{t-1}$to update its state $\mathbf{s}_t$ as $\mathbf{s}_t = f(\mathbf{s}_{t-1}, y_{t-1}, \mathbf{c}_t)$, and $\mathbf{c}_t$is a weighted average of the encoder hidden states as,

$$\begin{gathered} e_{ti} = g(\mathbf{s}_{t-1}, \mathbf{h}_i), \quad \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{T} \exp(e_{tk})}, \\ \mathbf{c}_t = \sum_{i=1}^{T} \alpha_{ti} \mathbf{h}_i, \end{gathered} \tag{1}$$

where g is inner-product function and α_{ti} can be viewed as the similarity score between encoder's state $\mathbf{h}_t$ and decoder's state $\mathbf{s}_{t-1}$. Finally, the probability distribution of candidate words at time-step t is calculated as,

$$p(y_t|\mathbf{y}_{<t}, \mathbf{x}) = softmax(MLP(\mathbf{s}_t, y_{t-1})).$$

The final objective is to maximize the (log) likelihood of $p(\mathbf{y}|\mathbf{x})$ for all timesteps as $p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T'} p(y_t|\mathbf{y}_{<t}, \mathbf{x})$.

This Seq2Seq model lacks the parametrization of stochastic variations of generation process, where only source of stochastic variation is provided by conditional distribution of $p(y_t|\mathbf{y}_{<t}, \mathbf{x})$(the softmax layer). This model generates high-frequency comments with low-diversity [Li *et al.*, 2016].

3 Methodology

To improve comment diversity, we reveal inherent multi-level hierarchy in comment generation, and propose a probabilistic model that exploits latent variables and a stochastic attention.

Hierarchical structure. As Figure 2, first, a decision-level latent variable $\mathbf{z}_d$ is used to characterize the choice of high-level properties, such as topic, intention. Second, a series of word-level latent variables $\mathbf{z}_w = \{\mathbf{z}_w^t\}_{t=1}^{T'}$ are introduced to characterize the variations of word-level selections, generate diverse comments by focusing on different input key-words.

To induce a two-level hierarchy, we endow the latent variables $\mathbf{z}_d$ and $\mathbf{z}_w$ with structured dependencies by defining,

$$p(\mathbf{y}, \mathbf{z}_w, \mathbf{z}_d|\mathbf{x}) = p(\mathbf{y}, \mathbf{z}_w|\mathbf{z}_d, \mathbf{x}) p(\mathbf{z}_d|\mathbf{x}).$$

$$p(\mathbf{y}, \mathbf{z}_w|\mathbf{z}_d, \mathbf{x}) = \prod_{t=1}^{T'} p(y_t|\mathbf{z}_w^{\leq t}, \mathbf{y}_{<t}, \mathbf{z}_d, \mathbf{x}) p(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{<t}, \mathbf{z}_d)$$

i) discourse-level stochastic variations are modeled by the prior distribution of $p(\mathbf{z}_d|\mathbf{x})$. *ii*) word-level stochastic variations are modeled by the distribution of $p(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{<t}, \mathbf{z}_d)$. Specifically, each word-level variable $\mathbf{z}_w^t$ is not only conditioned on $\mathbf{x}$ and $\mathbf{z}_d$, but also conditioned on the previous latent trajectories of $\mathbf{z}_w^{<t}$. These structured dependencies promote the model capacity for capturing multi-level diversities. Rather than directly maximizing $p(\mathbf{y}|\mathbf{x})$, we define the joint probability distribution $p(\mathbf{y}, \mathbf{z}_w, \mathbf{z}_d|\mathbf{x})$ and maximize $p(\mathbf{y}|\mathbf{x})$ by marginalizing out all the latent variables as,

$$p(\mathbf{y}|\mathbf{x}) = \int_{\mathbf{z}_d, \mathbf{z}_w} p(\mathbf{y}, \mathbf{z}_w, \mathbf{z}_d|\mathbf{x}) \ d\mathbf{z}_w d\mathbf{z}_d. \tag{2}$$

Stochastic attention. One key hypothesis is that diverse comments can be generated by focusing on different key-words of input. Different from deterministic Soft-attention (Eq. 1), we explicitly inject stochastic variations in the context vector $\mathbf{c}_t$, by leveraging the word-level variables $\mathbf{z}_w^t$ as,

$$\begin{gathered} e_{ti} = g(\mathbf{z}_w^t, \mathbf{h}_i), \quad \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{T} \exp(e_{tk})}, \\ \mathbf{c}_t = \sum_{i=1}^{T} \alpha_{ti} \mathbf{h}_i, \end{gathered} \tag{3}$$

where the variations of the focused input-words are characterized by the latent distribution of $\mathbf{z}_w$. Stochastic attention helps better capture the 1-to-n mappings. For one input, more diverse input-words can be focused to generate informative and meaningful predictions.

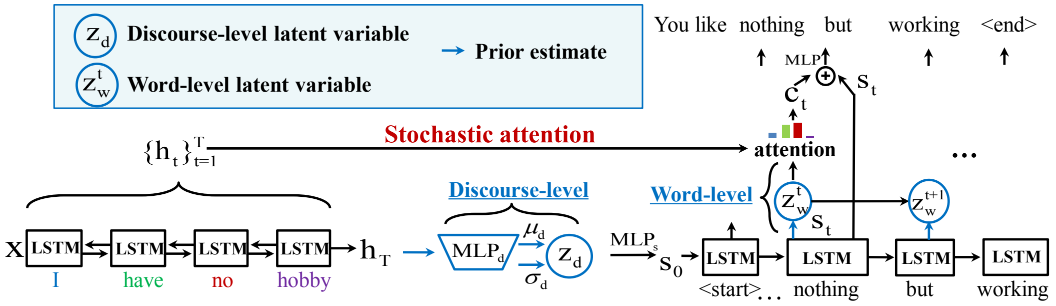


Figure 2: The network architecture for comment generation. Some of the MLP subnetworks are omitted for abbreviation.

Generation. As Figure 2, given an input $\mathbf{x}$, a comment sentence $\mathbf{y}$ is generated as follows.
1) Calculate the encoding states $\{\mathbf{h}_t\}_{t=1}^T$using a B-RNN encoder (see preliminary).
2) Sample a discourse-level variable $\mathbf{z}_d$ from a multivariate Gaussian distribution $p(\mathbf{z}_d|\mathbf{x})$as,

$$\mathbf{z}_d \sim p(\mathbf{z}_d|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_d, diag(\boldsymbol{\sigma}_d^2)). \tag{4}$$

The mean $\boldsymbol{\mu}_d$ and variance $\boldsymbol{\sigma}_d^2$ are produced by a MLP network, which takes the last hidden states $\mathbf{h}_T$as input as,

$$[\boldsymbol{\mu}_d, \boldsymbol{\sigma}_d] = MLP_d(\mathbf{h}_T). \tag{5}$$

3) Iteratively run step 4), 5), 6) for each step $t = 1, 2, ..., T'$.
4) Update the hidden state $\mathbf{s}_t$ of the decoder RNN as,

$$\mathbf{s}_t = f(\mathbf{s}_{t-1}, y_{t-1}, \mathbf{z}_w^{t-1}), \tag{6}$$

where f is a LSTM unit, y_0 is the '<start>' token, and $\mathbf{z}_w^0$ and $\mathbf{s}_0$ are initialized as $\mathbf{z}_w^0 = \mathbf{z}_d$ and $\mathbf{s}_0 = MLP_s(\mathbf{z}_d)$.
5) Stochastic-attention: Sample the t-th variable $\mathbf{z}_w^t$ from another Gaussian distribution $p(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{<t}, \mathbf{z}_d)$with mean $\boldsymbol{\mu}_t$and diagonal covariance $diag(\boldsymbol{\sigma}_t^2)$as

$$\mathbf{z}_w^t \sim p(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{<t}, \mathbf{z}_d) = \mathcal{N}(\boldsymbol{\mu}_t, diag(\boldsymbol{\sigma}_t^2)), \tag{7}$$

where the mean and variance are also produced by a MLP network with input $\mathbf{s}_t$ as $[\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t] = MLP_w(\mathbf{s}_t)$. Then, compute the context vector $\mathbf{c}_t$ according to Eq. 3.
6) Sample a word y_t with probability $p(y_t|\mathbf{y}_{<t}, \mathbf{z}_w^{\leq t}, \mathbf{z}_d, \mathbf{x})$ from vocabulary set V as,

$$\mathbf{p}_{y_t} = softmax(MLP(\mathbf{s}_t, \mathbf{z}_w^t, \mathbf{c}_t)) \in \mathbb{R}^{|V|}. \tag{8}$$

Inference. Directly optimizing objective of Eq. 2 involves a marginalization over the latent variables of $\mathbf{z}_d$ and $\mathbf{z}_w$, which introduces an intractable inference of the posterior $p(\mathbf{z}_d, \mathbf{z}_w|\mathbf{y}, \mathbf{x})$. Exploiting Stochastic Gradient Variational Bayes (SGVB) [Kingma and Welling, 2013], we introduce an auxiliary distribution $q(\mathbf{z}_d, \mathbf{z}_w|\mathbf{y}, \mathbf{x})$ to approximate the true posterior as,

$$q(\mathbf{z}_w, \mathbf{z}_d|\mathbf{y}, \mathbf{x}) = q(\mathbf{z}_d|\mathbf{y}, \mathbf{x}) \prod_{t=1}^{T'} q(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{\leq t}, \mathbf{z}_d),$$

where $q(\mathbf{z}_d|\mathbf{y}, \mathbf{x})$and $q(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{\leq t}, \mathbf{z}_d)$are multivariate Gaussian distributions with diagonal covariance, parameterized by MLP networks as,

$$\begin{aligned} q(\mathbf{z}_d|\mathbf{y}, \mathbf{x}) &= \mathcal{N}(\hat{\boldsymbol{\mu}}_d, diag(\hat{\boldsymbol{\sigma}}_d^2)), \\ [\hat{\boldsymbol{\mu}}_d, \hat{\boldsymbol{\sigma}}_d] &= MLP_{\hat{d}}(\mathbf{h}_T, \mathbf{h}_T^{\mathbf{y}}), and \end{aligned} \tag{9}$$

$$\begin{aligned} q(\mathbf{z}_w^t|\mathbf{z}_w^{<t}, \mathbf{y}_{\leq t}, \mathbf{z}_d) &= \mathcal{N}(\hat{\boldsymbol{\mu}}_t, diag(\hat{\boldsymbol{\sigma}}_t^2)), \\ [\hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\sigma}}_t] &= MLP_{\hat{w}}(\mathbf{s}_t, y_t). \end{aligned} \tag{10}$$

As shown in Figure 3, $\mathbf{h}_T^{\mathbf{y}}$ is the last hidden state of the B-RNN encoder with input $\mathbf{y}$. Via variational inference, we derive the objective function which is the lower bound of the conditional log likelihood of Eq. 2 as,

$$\begin{aligned} \mathcal{L} = \mathbb{E}_{q(\mathbf{z}_d, \mathbf{z}_w|\mathbf{y}, \mathbf{x})}[\sum_{t=1}^{T'} (log\ p(y_t|\mathbf{z}_w^{\leq t}, \mathbf{y}_{<t}, \mathbf{z}_d, \mathbf{x}) \\ -KL(q(\mathbf{z}_w^t|\mathbf{y}_{\leq t}, \mathbf{z}_w^{<t}, \mathbf{z}_d, \mathbf{x})||p(\mathbf{z}_w^t|\mathbf{y}_{<t}, \mathbf{z}_w^{<t}, \mathbf{z}_d)))] \\ -KL(q(\mathbf{z}_d|\mathbf{y}, \mathbf{x})||p(\mathbf{z}_d|\mathbf{x})) \quad \leq \quad log\ p(\mathbf{y}|\mathbf{x}), \end{aligned} \tag{11}$$

where KL is the Kullback-Leibler divergence, regularizing the approximated posteriors to be close to the priors.

Training. In Figure 3, the model is trained by maximizing the variational lower bound $\mathcal{L}$ of Eq. 11. In practice, the expectation term can be approximated using M Monte Carlo samples $\{\hat{\mathbf{z}}_d^{(m)}, \hat{\mathbf{z}}_w^{(m)}\}_{m=1}^M$ sampled from $q(\mathbf{z}_d, \mathbf{z}_w|\mathbf{y}, \mathbf{x})$. The reparametrization trick [Kingma and Welling, 2013] is adopted to reduce the variance of the approximation as,

$$\hat{\mathbf{z}}_d^{(m)} = \hat{\boldsymbol{\mu}}_d + \hat{\boldsymbol{\sigma}}_d \odot \boldsymbol{\varepsilon}^{(m)}, \ \ \hat{\mathbf{z}}_w^{t(m)} = \hat{\boldsymbol{\mu}}_t + \hat{\boldsymbol{\sigma}}_t \odot \boldsymbol{\varepsilon}^{(m)}, \tag{12}$$

where $\varepsilon^{(m)}$ is a vector of standard Gaussian samples, and $\odot$ is dot production. In practice, we set $M = 1$ and maximize the approximated variational lower bound $\hat{\mathcal{L}}$ as,

$$\begin{aligned} \hat{\mathcal{L}} = \tfrac{1}{M} \sum_{m=1}^{M} \sum_{t=1}^{T'} [\ log\ p(y_t|\mathbf{y}_{<t}, \hat{\mathbf{z}}_w^{\leq t(m)}, \hat{\mathbf{z}}_d^{(m)}, \mathbf{x}) - \\ KL(q(\mathbf{z}_w^t|\mathbf{y}_{\leq t}, \hat{\mathbf{z}}_w^{<t(m)}, \hat{\mathbf{z}}_d^{(m)}, \mathbf{x})||p(\mathbf{z}_w^t|\mathbf{y}_{<t}, \hat{\mathbf{z}}_w^{<t(m)}, \hat{\mathbf{z}}_d^{(m)}))] \\ -KL(q(\mathbf{z}_d|\mathbf{y}, \mathbf{x})||p(\mathbf{z}_d|\mathbf{x})) \quad \approx \quad \mathcal{L}. \end{aligned} \tag{13}$$

Analogous to existing latent variable models [Bowman *et al.*, 2015; Zhao *et al.*, 2017], our model suffers from the vanishing latent variable problem. During training, the KL terms

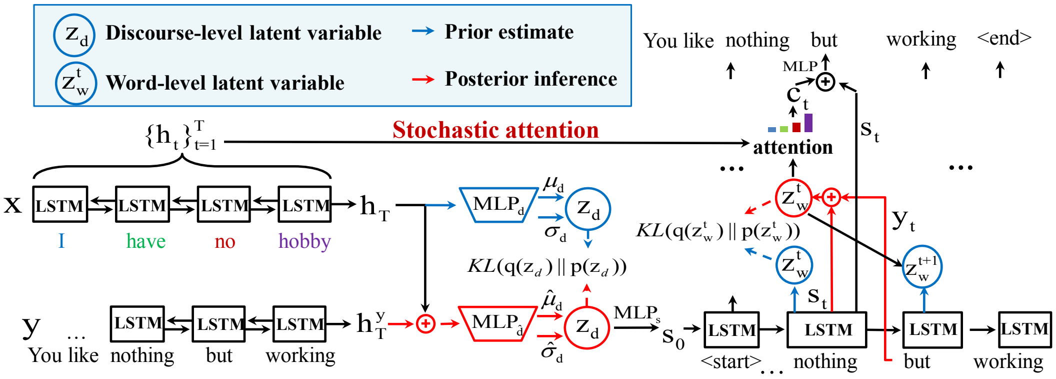


Figure 3: The network architecture for training. Some of the MLP subnetworks are omitted for abbreviation.

in Eq. 13 are easily crushed to zero, which means that latent variables fail to capture meaningful information. Since simply applying the KL-annealing [Bowman *et al.*, 2015] is unsatisfactory, we introduce extra strategies for $\mathbf{z}_d$ and $\mathbf{z}_w$.

1) For $\mathbf{z}_d$, the *bag-of-word* method [Zhao *et al.*, 2017] is adopted. An auxiliary objective $log\, p(\mathbf{y}_{bow}|\mathbf{z}_d)$ is used to force $\mathbf{z}_d$ to capture global bag-of-words information $\mathbf{y}_{bow}$ as,

$$\begin{aligned}\mathbf{p}_{y_b} &= softmax(MLP_{bow}(\mathbf{z}_d)),\\ log\, p(\mathbf{y}_{bow}|\mathbf{z}_d) &= \sum_{V_j \in \mathbf{y}_{bow}} log\, \mathbf{p}_{y_b}[j],\end{aligned} \tag{14}$$

where $\mathbf{p}_{y_b}$ is the probability vector over vocabulary V.

2) For $\mathbf{z}_w$, we propose an **Auxiliary-path method** with objective $log\ \tilde{p}(y_t|\mathbf{z}_w^t)$, which forces the model to make good predictions only conditioned on the latent variables $\mathbf{z}_w$. The auxiliary probability vector over the vocabulary for $\tilde{p}(y_t|\mathbf{z}_w^t)$ can be calculated by an auxiliary network MLP_{aux} as,

$$\tilde{\mathbf{p}}_{y_t} = softmax(MLP_{aux}(\mathbf{z}_w^t) \in \mathbb{R}^{|V|}. \tag{15}$$

The final objective function is rewritten as in Eq. 16, where α and β are hyper-parameters to balance two auxiliary terms.

$$\begin{aligned}\mathcal{L}_{final} &= \hat{\mathcal{L}} + \mathbb{E}_{q(\mathbf{z}_d,\mathbf{z}_w|\mathbf{y},\mathbf{x})}[\alpha\ log\, p(\mathbf{y}_{bow}|\mathbf{z}_d)\\ &+\beta \sum_{t=1}^{T'} log\ \tilde{p}(y_t|\mathbf{z}_w^t)],\end{aligned} \tag{16}$$

There are two differences between SARN and existing methods [Park *et al.*, 2018; Serban *et al.*, 2017; Shen *et al.*, 2017] which also exploit hierarchical structure. (1) they model different hierarchies, i.e., sentence-word hierarchy for SARN and dialog-sentence hierarchy for existing methods. (2) we leverage stochastic attention with hierarchical structure, which gives more capacity to model the complex variations in word selection.

4 Experiments

Data collection. We collected training corpus from Sina Weibo including Chinese post-comment data. To construct diverse training set, we compute scores of Tf-Idf cosine distance between a given post and other posts in the corpus, and then we select the nearest neighbor comments from the top-15 similar posts. We empirically find that these selected comments are coherent enough with the selected post. Finally, we build 1 million post-comments pairs, and each post is coupled with 15 different comments. The dataset is randomly split into training, validation and testing sets, with $980k$, $10k$ and $10k$ samples respectively.

Compared methods. 1) Seq2Seq [Sutskever *et al.*, 2014] Basic encoder-decoder model. 2) Seq2Seq-MMI [Li *et al.*, 2016] Seq2Seq model with Maximum Mutual Information objective function. We used the MMI-bidi version with parameters $\lambda = 0.5$, $\gamma = 1$. 3) VRNN [Chung *et al.*, 2015] Latent variable Seq2Seq model using time-step dependency. 4) CVRNN [Bowman *et al.*, 2015] Latent variable Seq2Seq model using global dependency. 5) SpaceFusion [Gao *et al.*, 2019] Seq2Seq model using latent space fusion[1].

Settings. We adopted the Jieba Chinese tokenizer [Jieba, 2018], and constructed a vocabulary with $40k$. For compared models, we adopted a LSTM for encoding and decoding. The dimension of the word embedding, attention state, latent variables, and LSTM hidden state were respectively set to be 100, 200, 500 and 500. The network parameters were initialized from normal distribution $\mathcal{N}(0, 0.01)$. We used Adam as optimizer with a fixed learning rate of 0.0001, and the batch size was set to 20. Gradient clipping was adopted with a threshold of 10. We selected parameters with the best validation performance. For the vanishing latent variable problem, we adopted the KL-annealing for models of VRNN, CVRNN and SARN, by linearly increasing training weight of KL-term from 0 to 1 after running 60,000 batches. We also adopted bag-of-word method for CVRNN, and the proposed auxiliary-path method for VRNN. The source code is released[2].

Evaluation methods. Given one post, we sample 5 comments for each model for evaluation. We adopt beam search in decoding process for Seq2Seq and Seq2Seq-MMI, where beam size was set to 10. For VRNN, CVRNN and SARN, we randomly sample 5 comments using greedy decoding (beam size is 1), where the randomness comes from the latent variables [Bowman *et al.*, 2015]. We apply three evaluation metrics, including human judgments, embedding-based evaluation, and n-gram based evaluation (BLEU-4). As [Zhou *et al.*, 2017a], human judgments is performed via crowd-sourcing.

[1] https://github.com/golsun/SpaceFusion

[2] https://github.com/ysjakking/weibo-comments

Models	%Acceptable	%Bad	%Normal	%Good	Diversity	BLEU-4	Average	Extrema	Greedy
Seq2Seq	61.4	38.6	44.7	16.7	1.51	0.413	60.75	83.15	60.35
Seq2Seq-MMI	65.8	34.2	43.3	22.5	1.87	**0.432**	59.08	82.64	58.45
VRNN	67.8	32.2	47.4	20.4	2.49	0.423	57.93	82.85	56.67
CVRNN	68.2	31.8	43.1	25.1	2.54	0.429	60.24	82.95	60.25
SpaceFusion	69.0	31.0	40.5	28.5	2.82	0.431	60.89	83.58	61.20
SARN	**69.3**	**30.7**	**40.1**	**29.2**	**2.95**	0.427	**61.05**	**83.60**	**61.21**

Table 1: The evaluation results of human judgements, BLEU-4 score and embedding-based score of Average, Extrema and Greedy.

Models	Wins	Losses	Ties
SARN vs Seq2Seq	55.6	11.3	33.1
SARN vs Seq2Seq-MMI	46.4	22.2	31.4
SARN vs VRNN	42.2	20.5	37.3
SARN vs CVRNN	36.1	23.7	40.2
SARN vs SpaceFusion	23.2	19.5	57.3

Table 2: The pairwise comparisons from human judgements.

α ($\beta = 0.05$)	$KL(\mathbf{z}_d)$	Diversity	Acceptable(Top-5)
0	0.1	1.53	61.4
0.1	2.2	2.31	63.2
0.3	5.6	2.95	**69.3**
0.7	7.9	3.12	65.2
1	10.2	3.10	62.5

Table 3: Parameter analysis for hyper-parameter α.

β ($\alpha = 0.3$)	$\Sigma KL(\mathbf{z}_t)$	Diversity	Acceptable(Top-5)
0	0.7	1.90	61.2
0.03	4.2	2.14	63.7
0.05	5.6	2.95	**69.3**
0.1	13.4	3.88	53.3
0.2	20.2	4.20	50.1

Table 4: Parameter analysis for hyper-parameter β.

We conduct blind evaluation, and outputs are randomly presented to 10 evaluators who have experiences of NLP. For each test input, every model generates 5 responses as a group. For each response the evaluators are asked to score its quality with Bad (NOT grammatically correct or relevant), Normal (grammatically correct and relevant to the input) and Good (beyond grammatically correct and relevant, the response is interesting and meaningful). Further, we define a comment is 'Acceptable' if it is scored Normal or Good. Besides, we conduct pairwise comparisons, where a better comment is chosen from two compared models. Numbers of Wins, Losses and Ties are reported. Compared with n-gram based metrics, embedding-based evaluation focuses more on measuring the semantic similarity. As [Serban *et al.*, 2017], three metrics are adopted, including Average, Extrema and Greedy.

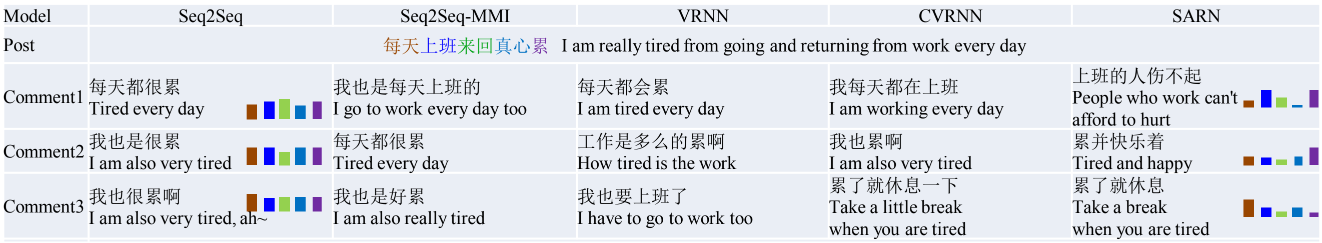
4.1 Performance Analysis

For human judgements, we evaluate the consistency of agreement using Fleiss' Kappa criterion, which are 0.34 (Fair agreement) for Table 1 and 0.30 (Fair agreement) for Table 2 respectively. We observe that, first, SARN outperforms other models with respect to both the Diversity value and Acceptable ratio. Compared with models only improving single-level diversity, such as the word-level VRNN and the discourse-level CVRNN, Seq2Seq-MMI and SpaceFusion, our method achieved higher Diversity value and Acceptable ratio. The improvements shows the effectiveness of leveraging the multi-level diversity. Second, the statistics of Bad, Normal and Good show that our model tends to generate more Good comments than Normal or Bad comments. It reveals that promoting the multi-level diversity also helps to improve the comment quality, generating more meaningful and interesting comments. Third, our method does not achieve superior BLEU-4 score. The reason is that BLEU-4 calculates the n-gram overlapping with ground-truth sentences, and such measurement is not positively related to the comment diversity. Because of such limitations of n-gram based measurement, we provide embedding-based evaluations. According to the Average, Extrema and Greedy scores in Table 1, we see SARN generates more coherent and informative comments in terms of high-level semantics. Finally, the pairwise comparison results in Table 2 show that the comments of our SARN are preferred in the majority of the experiments, and our method achieves comparable performance with SpaceFusion. It further validates that our SARN produces more diverse comments with better quality.

Figure 4 shows comment examples. We can see that SARN generates more informative comments with high diversity while other methods tend to produce phrases with similar meanings. As Eq. 3, comments are generated based on the understanding of key-words of interest. Color histogram in Figure 4 denotes the accumulated attention weights on the post word, where we find (1) SARN can focus on different post words so that it generates diverse comments; (2) Seq2Seq focuses on post words uniformly so that their comments are very similar. This is because that Seq2Seq uses deterministic soft-attention while SARN uses stochastic attention.

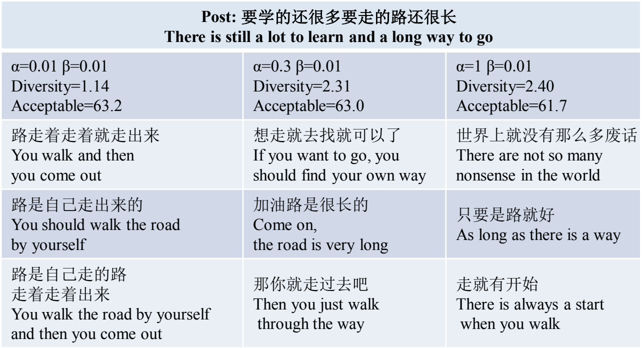
4.2 Parameter Analysis

To tackle the vanishing latent variable problem, we respectively add *bag-of-word* and *auxiliary-path* objectives for $\mathbf{z}_d$ and $\mathbf{z}_w$ in Eq. 16, introducing hyper-parameters α and β to balance the auxiliary terms. As shown in Table 3 and 4, we find that, first, $\mathbf{z}_d$ and $\mathbf{z}_w$ have great influence on Diversity degree and Acceptable rate. $\mathbf{z}_d$ is to characterize the global properties of sentencesand and $\mathbf{z}_w$ mainly controls the word-choice. Second, both of KL cost and Diversity are gradually increased with the increment of α and β. This indicates that the vanishing latent variable problem is alleviated. Third, Acceptable rate is first improved with the increase of α and β.



Model	Seq2Seq	Seq2Seq-MMI	VRNN	CVRNN	SARN
Post	每天上班来回真心累 I am really tired from going and returning from work every day				
Comment1	每天都很累 Tired every day	我也是每天上班的 I go to work every day too	每天都会累 I am tired every day	我每天都在上班 I am working every day	上班的人伤不起 People who work can't afford to hurt
Comment2	我也是很累 I am also very tired	每天都很累 Tired every day	工作是多么的累啊 How tired is the work	我也累啊 I am also very tired	累并快乐着 Tired and happy
Comment3	我也很累啊 I am also very tired, ah~	我也是好累 I am also really tired	我也要上班了 I have to go to work too	累了就休息一下 Take a little break when you are tired	累了就休息 Take a break when you are tired

Figure 4: Comment examples. The color histograms represent the accumulated attention weights on the post word of the same color.

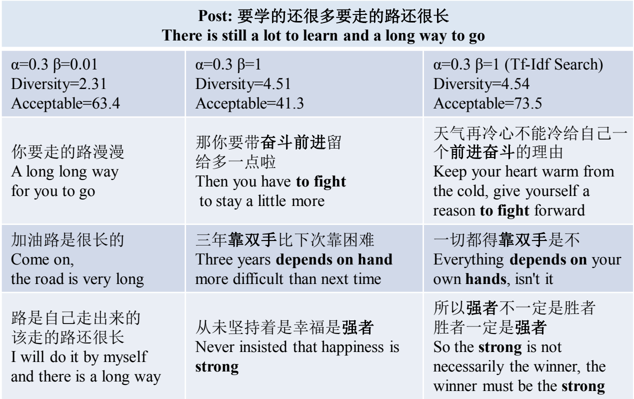


Post: 要学的还很多要走的路还很长 **There is still a lot to learn and a long way to go**		

α=0.01 β=0.01
Diversity=1.14
Acceptable=63.2
	α=0.3 β=0.01
	Diversity=2.31
	Acceptable=63.0
		α=1 β=0.01
		Diversity=2.40
		Acceptable=61.7

路走着走着就走出来 You walk and then you come out	想走就去找就可以了 If you want to go, you should find your own way	世界上就没有那么多废话 There are not so many nonsense in the world
路是自己走出来的 You should walk the road by yourself	加油路是很长的 Come on, the road is very long	只要是路就好 As long as there is a way
路是自己走的路 走着走着出来 You walk the road by yourself and then you come out	那你就走过去吧 Then you just walk through the way	走就有开始 There is always a start when you walk

Figure 5: Comment examples of different α.



Post: 要学的还很多要走的路还很长 **There is still a lot to learn and a long way to go**		

α=0.3 β=0.01
Diversity=2.31
Acceptable=63.4
	α=0.3 β=1
	Diversity=4.51
	Acceptable=41.3
		α=0.3 β=1 (Tf-Idf Search)
		Diversity=4.54
		Acceptable=73.5

你要走的路漫漫 A long long way for you to go	那你要带**奋斗前进**留 给多一点啦 Then you have **to fight** to stay a little more	天气再冷心不能冷给自己一个**前进奋斗**的理由 Keep your heart warm from the cold, give yourself a reason **to fight** forward
加油路是很长的 Come on, the road is very long	三年**靠双手**比下次靠困难 Three years **depends on hand** more difficult than next time	一切都得**靠双手**是不 Everything **depends on** your own **hands**, isn't it
路是自己走出来的 该走的路还很长 I will do it by myself and there is a long way	从未坚持着是幸福是**强者** Never insisted that happiness is **strong**	所以**强者**不一定是胜者 胜者一定是**强者** So the **strong** is not necessarily the winner, the winner must be the **strong**

Figure 6: Comments examples of different β and Tf-Idf search.

Then the performance degradation is showed when $\alpha \geq 0.3$ and $\beta \geq 0.05$ respectively. The main reason is that too large values of α and β overemphasize the auxiliary objectives, which gives the model much freedom to generate ungrammatical sentences. These results show that we can control the trade-off between improving diversity and maintaining grammatical sentence by using different values of α and β.

4.3 Multi-level Diversity Analysis

Parameter analysis shows that the values of α and β are proportional to the 'Diversity' degree. We further investigated how these two parameters influence the results. First, we tested different α by fixing $\beta = 0.01$ as shown in Figure 5. We can see that with the increase of α, the generated comments contain more different meanings, which indicates that the latent variable $\mathbf{z}_d$ successfully catches discourse-level variations. Second, we tested different β by fixing $\alpha = 0.3$. We found that the 'Diversity' degree of local words is more sensitive to β than α. As shown in Figure 6, our model begins to generate ungrammatical sentences consisting of diverse key words using large values of $\beta > 0.1$. This observation indicates that latent variables $\mathbf{z}_w$ capture the word-level variations, and large value of β makes $\mathbf{z}_w$ dominate the word-decisions, neglecting the long-term sequence dependency of LSTM hidden units. The above observations confirm our assumption that the two kinds of latent variables respectively catch discourse-level and word-level variations.

4.4 Large β And Search-based Model

Interestingly, we found the ungrammatical sentences caused by large β ($\beta > 0.1$) usually contain highly diverse but coherent key words with respect to the post (see bolded words in the middle column of Figure 6). Based on this observation, we built an effective search-based model based on these generated key words. In detail, we simply conducted Tf-Idf nearest search and selected the most similar comments from the training set. As shown in the right column of Figure 6, the obtained comments are highly coherent with the input post, and the final Diversity score and Acceptable ratio are further improved. Therefore, this search-based model provides us an effective way to combine generative and search-based generative model for practical industry use.

5 Conclusion

As a probabilistic model, the proposed SARN exploits both hierarchical-structured latent variables and the stochastic attention to promote multi-level diversity of comments. SARN is highly related to encoder-decoder models, while the main difference is that we inject multi-level stochastic variations in the generation process with both hierarchical and temporal dependencies. Experiments show that our model generates more diverse and realistic comments than other methods.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grand: 2018AAA0102003, in part by National Natural Science Foundation of China: 61771457, 61732007, 61672497, 61772494, 61931008, U1636214, 61622211, 61702491, and in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013.

References

[Bowman *et al.*, 2015] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *Computer Science*, 2015.

[Cao and Clark, 2017] Kris Cao and Stephen Clark. Latent variable dialogue models and their diversity. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 182–187, 2017.

[Chung *et al.*, 2015] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. *Computer Science*, 35(8):1340–1353, 2015.

[Eric and Manning, 2017] Mihail Eric and Christopher D Manning. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*, 2017.

[Gao *et al.*, 2019] Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. Jointly optimizing diversity and relevance in neural response generation. *NAACL*, 2019.

[Goddeau *et al.*, 1996] D Goddeau, H Meng, J Polifroni, and S Seneff. A form-based dialogue manager for spoken language applications. In *International Conference on Spoken Language, 1996*, pages 701–704 vol.2, 1996.

[Holtzman *et al.*, 2019] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[Jieba, 2018] Jieba. https://pypi.org/project/jieba/. 2018.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *Computer Science*, 2016.

[Liu *et al.*, 2018] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. *ACM MM*, 2018.

[Liu *et al.*, 2019] Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. Adaptive reconstruction network for weakly supervised referring expression grounding. *IEEE ICCV*, 2019.

[Mou *et al.*, 2016] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*, 2016.

[Park *et al.*, 2018] Yookoon Park, Jaemin Cho, and Gunhee Kim. A hierarchical latent structure for variational conversation modeling. *NAACL*, 2018.

[Serban *et al.*, 2017] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301, 2017.

[Shao *et al.*, 2017] Louis Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating long and diverse responses with neural conversation models. *arXiv preprint arXiv:1701.03185*, 2017.

[Shen *et al.*, 2017] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. A conditional variational framework for dialog generation. *arXiv preprint arXiv:1705.00316*, 2017.

[Shen *et al.*, 2019] Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. Towards generating long and coherent text with multi-level latent variable models. *ACL*, 2019.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. 4:3104–3112, 2014.

[Vijayakumar *et al.*, 2016] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[Xing *et al.*, 2017] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *AAAI*, volume 17, pages 3351–3357, 2017.

[Yang *et al.*, 2019] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: History-advantage sequence training for visual dialog. *IEEE ICCV*, 2019.

[Yao *et al.*, 2016] Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292*, 2016.

[Zha *et al.*, 2019] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE Trans. on PAMI*, 2019.

[Zhang *et al.*, 2018] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820, 2018.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv*, pages 654–664, 2017.

[Zhou *et al.*, 2017a] Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. Mechanism-aware neural machine for dialogue response generation. In *AAAI Conference on Artificial Intelligence, AAAI*, 2017.

[Zhou *et al.*, 2017b] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*, 2017.