A GAUSSIAN MIXTURE MODEL FOR DIALOGUE GENERATION WITH DYNAMIC PARAMETER SHARING STRATEGY

Qingqing Zhu[1]*, Pengfei Wu*[1]*, Zhouxing Tan*[1]*, Jiaxin Duan*[1]*, Fengyu Lu*[1] *, Junfei Liu*[1]

[1]School of Software and Microelectronics, Peking University, Beijing, China

ABSTRACT

Existing dialog models are trained with data in an encoder-decoder framework with the same parameters, ignoring the multinomial distribution nature in the dataset. In fact, model improvement and development commonly requires fine-grained modeling on individual data subsets. However, collecting a labeled fine-grained dialogue dataset often requires expert-level domain knowledge and therefore is difficult to scale in the real world. As we focus on better modeling multinomial data for dialog generation, we study an approach that combines the unsupervised clustering and generative model together with a GMM (Gaussian Mixture Model) based encoder-decoder framework. Specifically, our model samples from the prior and recognition distributions over the latent variables by a Gaussian mixture network and the latent layer with the capability to form multiple clusters. We also introduce knowledge distillation to guide and improve the clustering results. Finally, we use a dynamic parameter sharing strategy conditioned on different labels to train different decoders. Experimental results on a widely used dialogue dataset verify the effectiveness of the proposed method.

Index Terms— dialogue generation, cluster, GMM, knowledge distill, dynamic parameter sharing strategy

1. INTRODUCTION

With the rapid boom of social media on the Internet, people have become accustomed to having conversations on various websites. Therefore, the research of data-driven open domain dialogue systems has been promoted. These large amount of data usually have different categories (e.g.,mood, topic, or sentence function). Some previous studies have added these related category information to their dialog generation architectures. For example, endowing text-based dialogue generation systems with emotions is an active research area [1, 2, 3, 4, 5]. Liu et al. [1] focus on customer service dialogue with sentiment classification, which aims to assign a proper sentiment label to each utterance in a customer service dialogue. What's more, some papers link dialog generation tasks with different types of topics [6, 7, 8, 9]. Incorporating sentence function as another mode also shows improvements in the quality of generated responses [10]. Sentence function is an important linguistic feature indicating the communicative purpose in uttering a sentence. Gao et al. handle the dialogue conditioned on different sentence functions as separate tasks, and model agnostic meta-learning is applied to high-resource sentence function data [10]. However, these models usually require a large amount of annotation data for training. Nevertheless, in many practical applications, labeled data is very limited and the acquisition cost is high. As a consequence, it is becoming increasingly important to learn in an unsupervised manner.

In this paper, we introduce a novel cluster-driven dialogue model that combines the unsupervised clustering and generative model together with a GMM based encoder-decoder framework. Figure 1 shows an overview of our framework. In this framework, firstly, the model samples from the prior and recognition distributions over the latent variables by a Gaussian mixture network and the latent layer, which is able to form multiple clusters. In order to process the classification results better and establish the unity of prior and recognition network, we also extend the clustering process with knowledge distillation [11], which aims to transfer the knowledge of *teacher* network to a *student* model. We define the soft labels obtained from Gaussian mixture prior network as teacher labels, while labels generated by recognition distributions as students. Then we learn to cluster via teachers' soft-assignment scores. Finally, we use a dynamic parameter sharing strategy for different decoders conditioned on different labels. We conducted several experiments on the Weibo[1] dataset and found that our proposed neural architectures allow models to generate both longer and contextual responses compared to the literature baselines.

2. METHODOLOGY

2.1. Deep Gaussian Mixture Model

Generative model. First, two RNN encoders are first applied to encode the query and reply into vectors q and r . Then the model infers the prior distribution z_1 of the latent variable under the condition of the q through a matrix multiplication to the output of the feed-forward network, which defines the multivariate normal mean and variance. Similarly, the recognition network takes the concatenation of both r and q as input and converts them into latent variable target z_2.

[1]https://www.weibo.com/

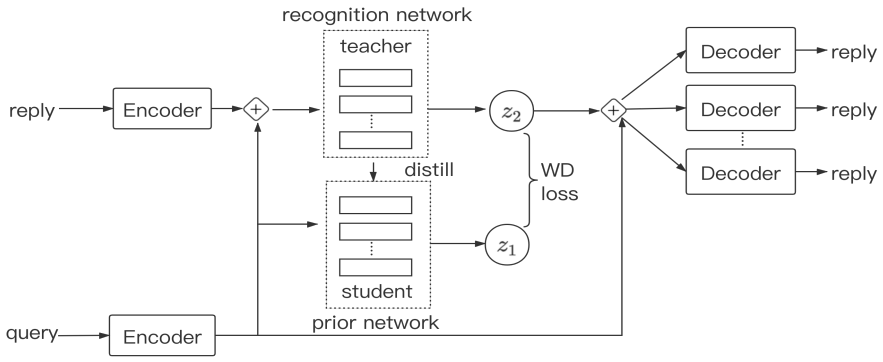

Fig. 1: The overview of our architecture. This is the train process of our approach. At the test time, we put z_1 into the decoder instead of z_2.

In order to measure the distance between the distribution $q_\phi(z_2|q,r)$ and $p_\theta(z_1|q)$, we adopt the Wasserstein distance as the nonparametric distance, and our objective function is:

$$L_D = W(p_\theta \| q_\phi), \tag{1}$$

where $W(\cdot\|\cdot)$ represents the Wasserstein distance. During the training time, we also maximize the log-probability of a reconstructed response from z_2, which is defined as:

$$L_R = E_{q_\phi(z_2|q,r)} \log p_\psi(r|z_2,q), \tag{2}$$

where $p_\psi(r|z_2,q)$ is a decoder.

For inference at test time, only q is available as input, and accordingly the latent variable z_1 that identifies recognition network cannot be used. Instead we generate a sample from the prior z_1, which is then used to compute reply.

Clustering model. In our model, we assume the data comes from multivariate Gaussian distribution. Specifically, we make the network capture a mixture of Gaussian distributions namely $\mathrm{GMM}\left(\left\{\pi_n\mu_n, \sigma_n^2 I\right\}_{n=1}^N\right)$ for estimating the probability of the category, where μ_n is the mean of the normal distribution for cluster n and σ_n is the covariance of the normal distribution for cluster n. Mixture coefficient π_n represents the degree of a node's association with cluster n. π is the distribution of the GMM. In our model, we obtain two probability distributions π^1,π^2 as the label distribution of prior and recognition networks, respectively.

2.2. Distill

Here our cluster procedure is built on the two classifiers of the prior and recognition network. We consider the classifiers of the recognition network to have the role of the teacher as it contains the information of both queries and replies, while the other network is the student.

We first apply a regularization method called entropy minimization through the use of a sharpening function [12]. It encourages the model to make low-entropy predictions on unlabeled data [13] by the following operation:

$$\pi_n = \mathrm{Sharpen}\left(\pi_n, T\right) = \sum_{n=1}^{N} \frac{(\pi_n)^{\frac{1}{T}}}{\sum_{n=1}^{N}(\pi_n)^{\frac{1}{T}}}, \tag{3}$$

where T is a hyperparameter that decides the temperature of entropy minimization. For all experiments we set $T = 0.5$. Both π_n^1 and π_n^2 are calculated in the same way, so we represent them by π_n in this equation.

Then, the teacher provides pseudo-labels as a target to refine student's predictions, we add this for the student network with the following objective:

$$L_{Distill} = \mathcal{H}(\pi^1, \pi^2), \tag{4}$$

where $\mathcal{H}$ refers to the cross-entropy. Then we have defined final training objective of our architecture accordingly:

$$L_G = L_R + L_D + L_{Distill}. \tag{5}$$

2.3. Decoders

Here we consider different decoders for different data according to their categories. Using completely independent parameterized decoders provides them with additional capacity to differentiate from one another. However, it may exacerbate overfitting. At the same time, some categories have less data, their decoders will be low-quality and neglected and eventually "die" during training. Therefore, we set up a dynamic sharing parameter process between decoders, which alleviates their homogenization [14]. Sharing parameters among decoders may help mitigate degeneracy, since by sharing parameters even unpopular decoders receive some gradients. We apply a gate $g(\cdot)$ to maintain the diversity of decoders [15]. During each epoch, we introduce a combined decoder and N individual decoders in our model. N is the number of clusters. Whether decoders share parameters with others subjects to $g(p) \sim$ Bernoulli (p). When $g(p)$ is 0, we first compute

the parameters of the nth individual decoder with the following function: $parameters_n = (1-p) * parameters_n + p * parameters_{combined}$, where $parameters_n$ are parameters of the nth individual decoder, $parameters_{combined}$ are the parameters of the combined decoder, then we use the related training corpus to update it. While when $g(p)$ equals to 1, the corpus of related categories is only put into the combined decoder for training. Besides, the gate also reduces computational cost. For example, we can keep the computational cost constant by adjusting the probability p when the number of clusters increases.

3. EXPERIMENT

Datasets. The assessment of our method is performed using a corpus obtained from Weibo. Specifically, the data comprises 400K query-reply pairs, which are made up of Weibo posts and their following replies. All these pairs are manually filtered with annotators by eliminating ungrammatical sentences and incoherent dialogues.

Automatic Evaluation Methods. In this paper, we adopt several evaluation methods to measure different aspects of our results: (1) BLEU [16]: it is used as a reward to evaluate dialog systems by measuring word overlap between the generated reply and the ground truth for the final evaluation. We compute BLEU scores for $n <= 4$ using smoothing techniques [2]. (2) BOW Embedding [17]: We adapt three embedding-based similarity metrics proposed by [17]: Greedy, Average and Extrema. (3) Entropy-based metrics: we compute the n-gram word entropy (n=1,2,3) as [18] and report them as Word Entropy-1,Word Entropy-2, Word Entropy-3, which suggests the non-genericness of responses. (4) Distinct [19]: we report the degree of diversity by calculating the proportion and number of distinct unigrams, bigrams and trigrams in generated responses. We represent them as Distinct-1, Distinct-2 and Distinct-3 in the table, respectively.

Human Evaluation Method. Considering the limitations of the existing automatic evaluation metrics, we also include human judgments. We randomly sample 100 cases and have recruited three well-educated volunteers to do manual evaluation. For each query-reply pair, volunteers are asked to rate it with three levels: 0, 1, 2. 0 indicates that the selected sentences are either irrelevant or disfluent with grammatical errors; 1 is for the reply that is relevant but not informative enough; 2 means that the queries and replies are extremely related and the replies are natural. We calculate the ratio of each score (0, 1 and 2) to evaluate the performance of each model. To estimate the agreements among all the volunteers, we also calculate the Fleiss Kappa [20] of the human annotations on all models.

Comparison Models. To ascertain the effectiveness and applicability of our approach, we re-implement experiments on these methods: (1) S2SA-MMI [19]: the Attn-Seq2Seq model which is decoded with maximum mutual information criterion.(2) CVAE: it is a latent variable model using conditional variational auto-encoder trained with KL annealing and a BoW loss as in [21].(3) GMoT-Seq2Seq [22]: a sequence to sequence model with a gated mixture of topics (MoT) designed to utilize topic information to generate coherent responses. (4) DialogWAE: a conditional Wasserstein auto-encoder that models the distribution of data by training a GAN within the latent variable space [23].(5) KD [24]: it uses two dialogue models as the student and the teacher. The teacher-student framework train student model by learning from both the ground-truth labels and the soft predictions of the teacher.

Training and Evaluation Details. We use gated recurrent units (GRU) for the RNN functions. The encoders and decoders are both GRUs with 256 hidden units. The prior and the recognition networks are both 2-layer feed-forward networks of size 200 with tanh activation function. The dimension of a latent variable z is set to 64. We sample the initial weights of all fully connected layers from the uniform distribution [-0.02, 0.02]. We set the vocabulary size to 40,000 and define all the out-of-vocabulary words to a special token $\langle unk \rangle$. The word embedding size is set to 200. These baselines are implemented using the same set of hyperparameters. All the models are implemented with Pytorch [3].

Automatic Evaluation Results. The automatic evaluation results of our proposed method and existing related methods are shown in Table 2. We can see the following observations: (1) Our model outperforms the baselines in terms of almost all the evaluation metrics, which means that our proposed model achieves better performance overall. (2) Especially, in terms of BLEU scores, compared to the S2S-MMI, CVAE, GMoT-Seq2Seq, DialogWae and KD, our model obtains impressive 20.34% , 16.45% , 8.18%, 12.23% and 8.04% performance gains. This indicates that our model produces more relevant responses with the highest BLEU scores on the dataset. (3) To demonstrate that our model is on average more diverse than other model responses, we compute the word entropy, and our model generates responses with higher entropy compared to the other baseline models.

Human Evaluation Results. Human evaluation results are presented in Table 1. The Kappa scores on all models are larger than 0.6, which indicates that the annotators had a fair level of agreement. From the results we can again observe that, similar to the automatic evaluation results, our model always achieves the best performance, which further proves the effectiveness of our proposed method.

Ablation Study. In the bottom of Table 1, we show the results of model variants by ablation the specific parts of our model. W/o distill means that we ablate the distill process from the framework, we can see that it hurts performance significantly on all evaluation metrics. Meanwhile, without

[2] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[3] https://pytorch.org/

Table 1: Results of the automatic evaluation and human evaluation (%) on the dataset. The metrics Average, Extrema and Greedy are abbreviated as A, E and G, respectively. The best results are highlighted with bold.

Method	BLEU	BOW Embedding			Distinct			Word Entropy			Human Evaluation			
		A	E	G	1	2	3	1	2	3	0	1	2	Kappa
S2SA-MMI	0.2188	0.590	0.526	0.605	0.448	0.514	0.545	4.514	5.667	7.081	35.8	49.2	15.0	0.639
CVAE	0.2261	0.597	0.530	0.609	0.473	0.538	0.562	4.634	5.901	7.435	27.8	53.2	19.0	0.653
GMoT-Seq2Seq	0.2434	0.527	0.468	0.531	0.500	0.570	0.597	4.717	6.17	7.613	28.7	43.9	27.4	0.625
Dialogwae	0.2346	0.480	0.426	0.479	0.495	0.570	0.596	4.709	6.072	7.424	26.3	49.0	24.7	0.693
KD	0.2437	0.533	0.473	0.544	0.466	0.527	0.551	4.556	5.766	7.108	27.2	45.0	27.8	0.692
Ours	**0.2633**	**0.636**	**0.565**	**0.649**	**0.521**	**0.595**	**0.623**	**4.786**	**6.267**	**7.753**	22.1	47.2	30.7	0.682
w/o distill	0.2447	0.537	0.478	0.546	0.469	0.529	0.553	4.560	5.740	7.047	-			
w/o sharpen	0.2511	0.634	0.561	0.647	0.527	0.601	0.628	4.784	6.263	7.696	-			

Table 2: Results of the automatic evaluation on the dataset.

p	N	BLEU	BOW Embedding			Distinct			Word Entropy		
			A	E	G	1	2	3	1	2	3
0	7	0.2525	0.480	0.553	0.568	0.512	0.578	0.603	4.713	6.078	7.352
0.3	7	0.2455	0.460	0.410	0.464	0.457	0.520	0.546	4.410	5.396	6.481
0.5	7	**0.2633**	**0.636**	**0.565**	**0.649**	**0.561**	**0.595**	**0.623**	**4.786**	**6.267**	**7.753**
0.7	7	0.243	0.500	0.527	0.468	0.531	0.570	0.597	4.717	6.170	7.613
1	7	0.2404	0.581	0.514	0.588	0.484	0.559	0.587	4.565	5.752	6.998

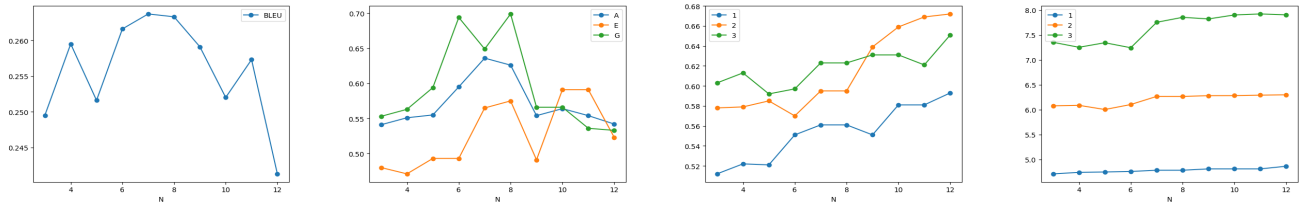

Fig. 2: Performance comparison with different number (N) of clusters on the dataset.

sharpen strategy also decreases the performance on most evaluation metrics, which further proves the effectiveness of combining these techniques together.

Impact of the Parameter Sharing Probability P. Table 2 presents the results of our model with the parameter sharing probabilities of 0, 0.3, 0.5, 0.7 and 1 by fixing $N = 7$. When the probability is 0, it means that the decoders don't share any parameters with each other. As an opposite, 1 means that only utilizing one combined decoder. We can see that when the parameter sharing probability is 0.5, our model achieves top performance. On the whole, the performance of our model first ascends and then declines as the parameter sharing probability increases. A plausible explanation is that utilizing independent decoders provides them with additional capacity to differentiate from one another, but exacerbates overfitting.

Impact of Different Number (N) of Clusters. By fixing $p = 0.5$, we vary N from 3 to 12 and plot the results in Figure 2. When N is set to 7 or 8, the quality of the responses generated by the model is better than others. Before N reaches 8, the score of BLEU and bow Embedding increases. While when $N > 8$, there was a marked decline in these two indicators. A plausible explanation is that it leads to the problem of overfitting in the case of too many groups. We also conjecture that marginal improvement in distinct and word entropy will be obtained by increasing the number of clusters further, which means that the diversity of generated responses increases. But in these cases, the value of BLEU and Bow Embedding drops a lot. Also, the cost of our model increases a lot. Considering these factors, we tend to set the number of clusters to 7 or 8 in the model.

4. CONCLUSION

In this work, we consider open-domain dialogue systems. To induce fine-grained model learning, we propose a method that combines the unsupervised clustering and generative model simultaneously. Also, based on the clustering results, we apply a dynamic parameter sharing strategy to train different decoders. Experimental results show that the proposed framework can effectively boost the performance of dialogue systems. Besides, our framework is not limited to the neural dialogue generation task. In the future, we would like to extend our method to deal with other text generation tasks like abstract summarization and examine it's adaptability.

5. REFERENCES

[1] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie

Huang, "Towards emotional support dialog systems," in *ACL/IJCNLP 2021*, pp. 3469–3483.

[2] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He, "Topic-driven and knowledge-aware transformer for dialogue emotion detection," in *ACL/IJCNLP 2021*, pp. 1571–1582.

[3] Young-Jun Lee and Ho-Jin Choi, "Comparative study of emotion annotation approaches in korean dialogue," in *BigComp 2021*, 2021, pp. 354–357.

[4] Deeksha Varshney, Asif Ekbal, and Pushpak Bhattacharyya, "Modelling context emotions using multi-task learning for emotion controlled dialog generation," in *EACL 2021, 2021*, pp. 2919–2931.

[5] Tatsuya Ide and Daisuke Kawahara, "Multi-task learning of generation and classification for emotion-aware dialogue response generation," in *NAACL-HLT 2021*, Esin Durmus, Vivek Gupta, Nelson Liu, Nanyun Peng, and Yu Su, Eds., pp. 119–125.

[6] Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu, "Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation," in *AAAI 2021*, pp. 14006–14014.

[7] Yi Xu, Hai Zhao, and Zhuosheng Zhang, "Topic-aware multi-turn dialogue modeling," in *AAAI 2021*, pp. 14176–14184.

[8] Yicheng Zou, Lujun Zhao, Yangyang Kang, Jun Lin, Minlong Peng, Zhuoren Jiang, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu, "Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling," in *AAAI 2021*, pp. 14665–14673.

[9] Karin Sevegnani, David M. Howcroft, Ioannis Konstas, and Verena Rieser, "Otters: One-turn topic transitions for open-domain dialogue," in *ACL/IJCNLP 2021*, pp. 2492–2504.

[10] Yifan Gao, Piji Li, Wei Bi, Xiaojiang Liu, Michael R. Lyu, and Irwin King, "Dialogue generation on infrequent sentence functions via structured meta-learning," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 431–440.

[11] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[12] Junnan Li, Richard Socher, and Steven C. H. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *ICLR 2020*.

[13] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *NeurIPS 2019*, 2019, pp. 5050–5060.

[14] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[15] Shaoxiong Feng, Xuancheng Ren, Kan Li, and Xu Sun, "Multi-view feature representation for dialogue generation with bidirectional distillation," in *AAAI 2021*. 2021, pp. 12812–12820, AAAI Press.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL 2002*, pp. 311–318.

[17] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *EMNLP 2016*, 2016.

[18] 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 2017*, 2017, pp. 3295–3301.

[19] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan, "A diversity-promoting objective function for neural conversation models," in *NAACL 2016*, 2016, pp. 110–119.

[20] Joseph L. Fleiss and Jacob Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational Psychological Measurement*, vol. 33, no. 3, pp. 613–619, 2016.

[21] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *ACL 2017*, pp. 654–664.

[22] Hongwei Zeng, Jun Liu, Meng Wang, and Bifan Wei, "A sequence to sequence model for dialogue generation with gated mixture of topics," *Neurocomputing*, vol. 437, pp. 282–288, 2021.

[23] Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim, "Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder," in *ICLR 2019,*.

[24] Amir Vakili Tahami, Kamyar Ghajar, and Azadeh Shakery, "Distilling knowledge for fast retrieval-based chatbots," in *SIGIR 2020*. pp. 2081–2084, ACM.