Variational Hierarchical User-based Conversation Model - Appendix

JinYeong Bak
School of Computing
KAIST
jy.bak@kaist.ac.kr

Alice Oh
School of Computing
KAIST
alice.oh@kaist.edu

Abstract

This document is an appendix of Variational Hierarchical User-based Conversation Model paper. It contains the description of our new Twitter conversation corpus and the parameter settings of VHUCM. It also has experiment results which are not shown or reduced size in the paper.

1 Twitter Conversation Corpus

This section describes our new conversation corpus from Twitter.

1.1 Building Twitter Conversation

At the starting point to crawl the tweets, we initialize the set of users by randomly sampling twenty-one users who reply to other users in English from the Twitter public streams named Gardenhose. Then we crawl each user's public tweets in his/her timeline and look at users who are mentioned in those tweets. It is a breadth-first search in the conversation network which is defined by users as nodes and edges as conversations. We run this search for conversation dyads until the depth of five, and filter out users who tweet in a non-English language by using an open source tool (García-Pablos et al., 2015).

We filter out dyads with fewer than ten conversations and the users who have less than three conversation friends. To protect users' privacy, we replace Twitter usernames and URLs in tweets with random strings.

Table 1 shows the basic statistics of our new corpus and other conversation corpora.

1.2 Getting the Corpus

We open the new Twitter conversation data in public. However, due to the privacy issues, we cannot open the utterance text directly even we anonymize the user personal information such as usernames and URLs. So, we decide to open the partial data used in new user experiments in the paper. We upload the data in GitHub repository[1]. Please get in touch with the first author if you want to access the data more.

2 VHUCM

This section describes VHUCM, and experiment hyperparameter setting to replicate the experiments results.

2.1 Parameter Settings

The RNN of VHUCM such as encoder, decoder, and context are GRU (Bahdanau et al., 2014). We use the pre-trained fastText word embedding (Mikolov et al., 2018) in encoder and decoder. We build user embedding vector from the conversation network by node2vec implementation[2] for VHUCM-PUE. We remove the words that appear less than five and set the size of the vocabulary is 20,000. We set 1,000 GRU hidden size, 300 speaker embedding size, 200 $\mathbf{z}_t^{utt}$ and $\mathbf{z}^{conv}$ size. The dropout ratio is 0.2 during the training time. We use Adam optimizer (Kingma and Ba, 2014) with 0.0001 learning rate. To solve vanishing latent variable problem, we adopt KL annealing where the KL multiplier increases from zero to one over 25,000 steps and add bag-of-word loss (Bowman et al., 2016; Zhao et al., 2017).

2.2 Implementation

We implement VHUCM using PyTorch. The url of the code is in the same GitHub repository of new corpus[1]. Please read the description in the repository to run the code.

Datasets	# Convs	# Utterances	# Dyads	# Users	*Avg utters*	*Avg convs*
New Twitter Conversation Corpus	770,739	6,109,469	107,611	27,152	7.92	7.16
Cornell movie corpus (Danescu-Niculescu-Mizil and Lee, 2011)	220,579	304,713	10,292	9,035	1.38	21.43
TV series transcripts (Li et al., 2016)	69,565	208,695	–	13	3.00	–
Ubuntu Dialog Corpus (Lowe et al., 2015)	930,000	7,100,000	–	–	3.84	–
Twitter conversation (Ritter et al., 2011)	≈ 1.3M	≈ 3M	–	–	≈ 2.31	–
Twitter triple conversation (Li et al., 2016)	≈ 29M	≈ 87M	–	74,003	3.00	–
Persona converation1 (Zhang et al., 2018)	≈ 700M	≈ 1,400M	–	≈ 5M	2.00	–
Persona converation2 Mazare et al. (2018)	13,201	161,898	10,970	1155	12.26	1.20
Dailydialog (Li et al., 2017)	13,118	102,980	–	–	7.85	–

Table 1: Basic statistics of our new Twitter conversation corpus and others. *Avg utters* is the average utterances in a conversation, and *Avg convs* is the average conversations in a dyad. M means million i.e., 3M = 3,000,000. ≈ means approximated value from the paper which describes the corpus, and – means hard to compute the value since the data is not opened or no information in the data.

3 Results of Response Quality

We use several automatic evaluation metrics to evaluate the generated responses from the models. The input context of the test data is three in the paper. Here, we add the results of one turn context since it is similar to the question and answering (QA). The general conversation is not the same as QA, but we show the user consistent answers in the paper. That's why we also show the quantitative results of the one-turn context case in this document. Table 2 shows that VHUCM-PUE outperforms all other models compared.

4 Examples of Generated Responses from Personal Questions

Table 3 shows the examples of the questions and responses. Additionally, we get the results to ask the "What did you have for dinner?" to users. Interestingly, user C uses ':)' and 'xx' words so we can imagine that they are close each other. And, A & D dyad are less close since A reveals the menu of the food to all users except D. These phenomenon also supports the discussion of the results in the paper.

[1] https://github.com/NoSyu/VHUCM

[2] http://snap.stanford.edu/node2vec

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the CoNLL*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the CMCL*.

Aitor García-Pablos, Montse Cuadros, and Maria Teresa Linaza. 2015. Opener: Open tools to perform natural language processing on accommodation reviews. In *Information and Communication Technologies in Tourism*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Model	BLEU	Embedding			ROUGE-L			Distinct		Len
		Avg	Ext	Gre	Rec	Prec	F	dist-1	dist-2	
HRED	0.061	0.532	0.338	0.334	0.061	0.158	0.062	0.014	0.036	5.3
VHRED	0.095	0.555	0.344	0.357	0.069	0.158	0.068	0.010	0.027	7.5
VHCR	0.107	0.566	0.347	0.370	0.073	0.166	0.072	0.013	0.035	**8.6**
SpeakAddr	0.047	0.543	**0.372**	0.333	0.049	**0.194**	0.052	0.001	0.000	4.9
DialogWAE	0.110	0.498	0.314	0.324	0.063	0.113	0.059	0.014	0.110	8.5
VHUCM	0.117	0.584	0.351	0.367	0.069	0.144	0.073	0.051	0.109	7.6
VHUCM-PUE	**0.128**	**0.591**	0.355	**0.373**	**0.077**	0.160	**0.080**	**0.067**	**0.149**	8.2

(a) Results of automatic evaluation metrics on 1-turn context input

Model	BLEU	Embedding			ROUGE-L			Distinct		Len
		Avg	Ext	Gre	Rec	Prec	F	dist-1	dist-2	
HRED	0.090	0.577	0.364	0.357	0.064	0.162	0.066	0.019	0.072	9.4
VHRED	0.120	0.596	0.368	0.377	0.072	0.161	0.072	0.016	0.063	11.4
VHCR	0.137	0.599	0.371	0.381	0.076	0.169	0.075	0.020	0.076	**12.3**
SpeakAddr	0.037	0.567	**0.384**	0.337	0.052	**0.218**	0.055	0.016	0.031	4.8
DialogWAE	0.127	0.586	0.345	0.369	0.079	0.132	0.080	0.012	0.104	11.5
VHUCM	0.120	0.633	0.373	0.394	0.075	0.154	0.079	0.030	0.108	10.1
VHUCM-PUE	**0.161**	**0.643**	0.376	**0.400**	**0.082**	0.162	**0.087**	**0.034**	**0.123**	10.6

(b) Results of automatic evaluation metrics on 3-turns context input

Table 2: Response quality test by automatic evaluation metrics given 1-turn and 3-turn utterances. BLEU, Embedding, and ROUGE-L compare generated response with ground truth. Embedding Avg, Ext, and Gre are average, extrema and greedy matching by embedding based metrics that pre-trained word vector embedding, respectively. ROUGE-L is ROUGE score by comparing the longest common subsequence between generated response and ground truth. Distinct is the degree of diversity among all generated response and is measured by the proportion of unique unigram (dist-1) and bigram (dist-2). Len is the average length of the generated response. Higher is better. VHUCM-PUE outperforms all other methods compared in terms of almost metrics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the ACL*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the IJCNLP*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL*.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the EMNLP*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the LREC*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the EMNLP*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the ACL*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the ACL*.

Questioner	Answerer	Questions		
		where is your hometown ?	how old are you ?	what did you have for dinner ?
User B	User A	north carolina !	i’m not sure , but i am a bit older than you	chicken and rice .
User C	User A	north carolina .	19 ! ! !	lamb , lamb and lamb .
User D	User A	north carolina .	i’m 19 . i don’t even know what to say	i don’t know , i don’t know .
User E	User A	north carolina .	i’m older than you .	i had a chicken salad .
User F	User A	north carolina ...	i dunno ...	chicken and rice .
User G	User A	north carolina .	19 . you ?	broccoli and broccoli
User A	User B	minnesota . <unk>.	18 yr old	steak ! ! ! :))
User A	User C	manchester :) xx	nothing much :)	it was yummy ! ! ! i had chicken and rice for dinner and it was lush :) xx
User A	User D	i live in <unk>.	i have no idea	lamb , lamb , lamb , <unk>
User A	User E	i am .	i don’t know .	<unk>! ! !
User A	User F	i live in texas	22 lol	chicken and waffles
User A	User G	st . louis .	i am old lol	steak and chips

Table 3: Responses of users’ personal information questions from VHUCM-PUE. The questioner ask the each question to the answerer, and VHUCM-PUE generates the answerers’ response. ‘<unk>’ token is unknown word. VHUCM-PUE generates consistent responses from the user.