Improved Multi-Round Chatbot using Data Augmentations and Loss Regularization

Meiwei Zhang
Athlone Institute of Technology
Athlone, Ireland
zhangmwchris@gmail.com

Yichang Wu
Athlone Institute of Technology
Athlone, Ireland
w.yichang@research.ait.ie

Xingzhen Ji
Brunel University London
London UK
xingzhenji@gmail.com

Jianxiang Gao
Brunel University London
London UK
Rexgao9586@hotmail.com

***Abstract*—A chatbot, interacting with users using natural language, is evolving towards a multi-round chatbot, which is more human-like. However, it suffers from the disability of changing the conversation subject smoothly. Since Knowledge Graph and Named Entity Recognition (NER) are widely practiced in Natural Language Processing (NLP) scenarios, Lian proposes a model which utilizes Knowledge Graph and NER to generate appropriate coherent response under specific knowledge [1]. Based on this model, we incorporate data augmentations and loss regularization methods to boost the human-computer dialogue performance. The augmentation methods consist of Entity Generation, Knowledge Matching, Dialogue Extraction. Our experiments witness an obvious improvement resulted from Entity Generation and Knowledge Matching. In particular, Knowledge Matching helps Encoder to exclusively select the highly relevant knowledge from both utterance and response, thus it greatly increases the word generation accuracy in decoder. Meanwhile, Loss Regularization is proved effective, due to avoiding over-fitting.**

***Index Terms*—Chatbot, Knowledge Graph, Seq2seq Model, Text Generation, Text Classification**

I. Introduction

Since iPhone 4S integrated built-in Siri, now various smart speaker applications have pervaded in our life, such as chatbot, recommendation system. All of them benefit from human-machine conversation technology which is one of the most important topics in artificial intelligence (AI). Such technology has received much attention across academia and industry in recent years. Currently dialogue system is still in its infancy. The chatbots usually converse passively and utter their words more as a matter of response rather than on their own initiatives.

Building a human-like conversational mechanism is one long-cherished goal in Artificial Intelligence (AI) [2]. Driven by dialogue corpus [3, 4], various kinds of conversational algorithms have been proposed during the past years, from the handcrafted rule-based systems [5] to the generation-based and retrieval-based systems. Although great progress has been made, Song concludes that generation-based approach tends to produce non-informative or universal response [6]. Meanwhile, most existing chatbot applications are suffering from the disability of changing the conversation subject smoothly. Furthermore, their response to human are logically chaotic rather than like the human-human conversations. It is difficult for agents to capture the connection between the previous and next conversations. To tackle the discrepancy, Lian proposes a novel knowledge selection mechanism where the knowledge is inferred from both the utterance and the response [1]. It proves that the model can effectively switch conversation subjects.

The innovation of this paper is combination of this model with data augmentations which are borrowed from computer version [7]. We propose three augmentation methods: Entity Generation is designed to increase the generalization ability. Knowledge Matching helps the conversation model to learn more closely related knowledge between subjects. Dialogue Extraction is adopted to increase example numbers.

The structure of the paper is as follows: In Section 2 we introduce the background of conversational mechanism. The model of our proposal is detailed in Section 3. We implement the model and analysis our experiments in Section 4. Finally, Section 5 conclude the paper.

II. Background

Conversational mechanisms experience a shift from handcrafted rule-based systems to generation-based systems. Early research has focused on extraction methods, which extract answer from database or prior pairs, and depend on frequency and stochastic topic models [8]. On the contrary, generation-based mechanism adopts Word Embedding technique as a basis. Word Embedding method converts a word as a continuous vector which captures the lexical and semantic characteristics of words in a dimensional space [9]. Our work is most directly related to continuous vector. For most NLP tasks, Blacoe et al. define each word as a dimension fixed vector which can be static or dynamic, and then process word embedding [10].

Generation-based conversational mechanism experiences great progress recently. With the rapid development of Deep Learning technology over the years, text generation method has been evolving. RNN-based structure is successfully applied to the area of NLP. Nallapati et al.

created encoder-decoder structure and attention mechanism to address generation tasks and obtained agreeable results in 2015[11]. Later, researchers extend the work and combine other features into the model to get better results. For example, graph-based attention mechanism proposed by Tan and Wan improves the generalization of the model [12]. Li et al. add history proposal as the latent variable into encoder to extract complex substructure [13]. Nallapati et al. use replication mechanism [11] to solve Out-Of-Vocabular (OOV) problems [14]. Mikolov et al. propose a hierarchical recurrent encoder-decoder neural network to build conversation mechanism [9]. Song et al. [6] design multi Seq2Seq models as generation chatbot modules. Both of these systems are designed without involving external knowledge. Recently Baidu NLP group, who has proved that, the knowledge-based conversation system, in general, can produce produce more accurate and informative response [1]. Moreover, Ghazvininejad et al. [15] adjust the Seq2Seq approach to a knowledge-grounded neural conversation model, which is able to response on the context of both conversation history and related knowledge.

III. Model

In this paper, the model is expected to produce proper replies based on history conversation and related knowledge, along with the topic goals. the conversation follows " START - $> TOPIC_A$ - $>TOPIC_B$". Formally, given a input sequence X = x_1,x_2,...,x_t, and a set of topic-related background knowledge K = k_1,k_2,...,k_n. and the corresponding response sequence Y_t as outputs, the model is aimed at learning the conditional probability distribution P(y|x,), which can be written as:

$$P(y|x,\Theta) = \prod_{t=1}^{t} P(y_i|y_{<i}, x_0, ...x_n, \Theta) \quad (1)$$

where the θ represents learnable parameters.

A. Encoder and Decoder

Chung et al. [16] use the encoder, a bidirectional RNN with a Gated Recurrent Unit (GRU) as shown in Fig. 1. It obtains the hidden state $\overrightarrow{h}_t$ for each x_t, $\overleftarrow{h}_t$ is the reverse order hidden state of x_t. these two hidden states are concatenated to from an overall hidden state $\boldsymbol{h_t}$ for x_t:

$$\boldsymbol{H_t} = [\overrightarrow{h}_t; \overleftarrow{h}_t] = [GRU(x_t, \overrightarrow{h_{t-1}}); GRU(x_t, \overleftarrow{h_{t+1}})] \quad (2)$$

where [;] represents a vector concatenation. Then this vector will be fed into the knowledge manager to facilitate knowledge selection and it will also serve as the initial hidden state of the decoder.

The decoder generates response word by word sequentially. It is assumed s_{t-1} is the last hidden state of the decoder, y_{t-1} is the word generated in the last step and c_t is an attention-based context vector of the encoder. The hidden state of the decoder at time t is:

$$s_t = GRU([y_{t-1}; k_i], s_{t-1}, c_t)c_t = \sum \alpha_{t,i} h_i \quad (3)$$

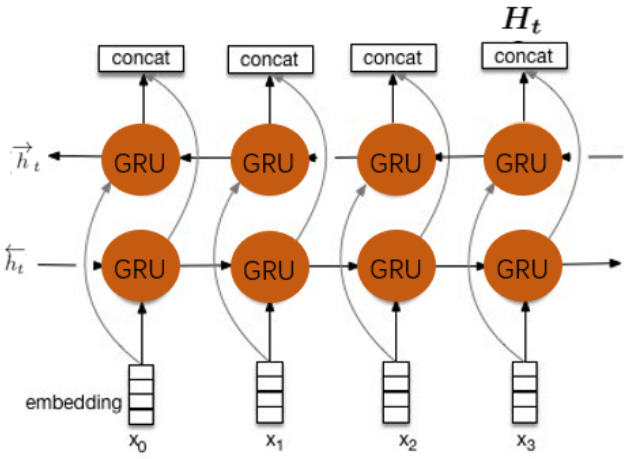

Fig. 1. Bidirectional GRU

where y_{t-1} is concatenated with k_i, which is the selected knowledge. $\alpha_{t,i}$ measures the relevancy between s_{t-1} and the hidden state h_i of the encoder. After obtaining the hidden state s_t, the next word y_t is generated according to the following correlation:

$$y_t \sim p_t = softmax(s_t, c_t) \quad (4)$$

B. Knowledge Manager

Fig 2 illustrates the details and functions of Knowledge Manager.

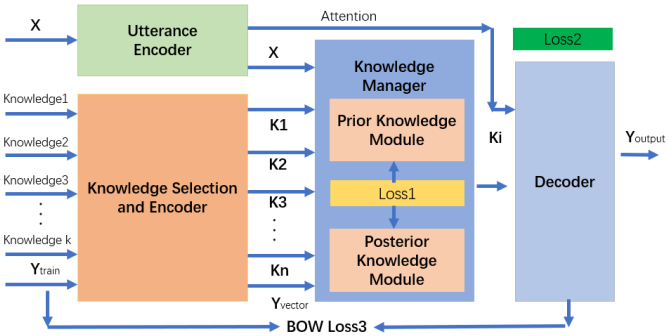

Fig. 2. Basic Structure with Knowledge Manager

The representations of conversations X, knowledge K and true response Y_{train} are fed to the Knowledge Manager. The responsibility of Knowledge Manager is to select related semantic background knowledge which leads the conversation naturally. In the knowledge module, we define a conditional probability distribution over knowledge, denoted by P(k|x) [17]:

$$p(k = k_i|x) = \frac{exp(k_i * x)}{\sum exp(k_j * x)} \quad (5)$$

the true knowledge reasoning distribution P(k_j|X,Y) [18]:

$$P(k_j|X, Y) = \frac{exp(k_j * MLP([x; y]))}{\sum exp(k_i * MLP([x; y]))} \quad (6)$$

$Loss_1$ is defined to evaluate the divergence between knowledge and response like human does. [1]:

$$loss_1 = \frac{1}{N}\sum P(k_j|X,Y)log\frac{P(k_j|X,Y)}{P(k_j|X)} \quad (7)$$

the difference between the true response and the response generated by the model. It minimizes the negative log-likelihood:

$$loss_2 = -\frac{1}{m}\sum P_\Theta(Y_t|Y_{<t},X,K) \quad (8)$$

To ensure the accuracy of output under certain knowledge, $loss_3$ depicts cross entropy by target output and decoder output. [19].

$$loss_3 = L_{BOW}(\Theta) = -\frac{1}{m}\sum logP(MLP(k_i)) \quad (9)$$

C. Loss Regularizer

Given a loss function $\sum_{i=1}^{n} V(f(x_i), y_i)$, a regularizer, $R(f)$, is added to it:

$$\min_f[\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)] \quad (10)$$

where V describes the cost of predicting $f(x)$ when the label is y, such as the square loss or hinge loss; and λ is a parameter which controls the importance of the regularization term. $R(f)$ is typically chosen to reduce the complexity of f. In this paper, we define Our loss regularization:

$$Loss(\theta) = Loss_1(\theta)*k_1 + Loss_2(\theta)*k_2 + Loss_3(\theta)*k_3 \quad (11)$$

where k_1, k_2, k_3 is our regularization parameters.

After experiments, we figure out the value of $Loss_3(\theta)$ is 500 times larger than the value of $Loss_2(\theta)$, and $Loss_1(\theta)$ is 1000 times larger than $Loss_2(\theta)$. In order to regularise them, k_1 is set as 0.001, k_2 is 1 and k_3 is 0.002.

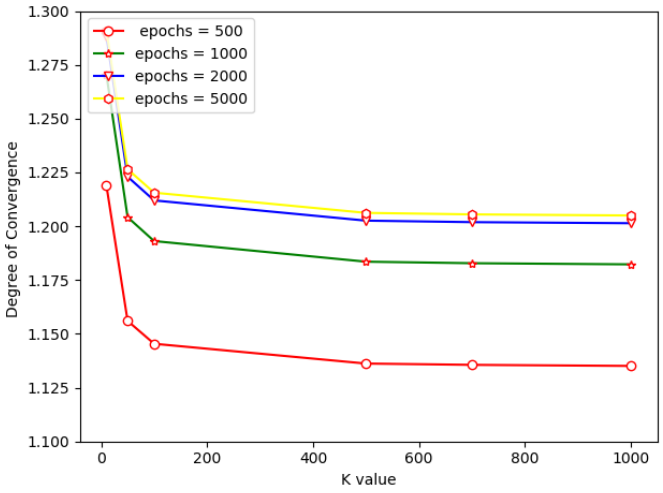

Fig. 3. K value correlation with degree of convergence

To verify this idea, we did an extra experiment based on liner regression. The loss function is given below:

$$Loss = Loss_{MSE} + k * Loss_{ABS} \quad (12)$$

we define $Loss_{MSE}$ is mean square error, $Loss_{ABS}$ is absolute error. The average value of $Loss_{MSE}$ and $Loss_{ABS}$ are in the same order of magnitude.

As k varies from 50 to 1000, degree of convergence, radio of initial loss value to end loss value, decreases as shown in Fig 3. From which we conclude:

- As k raising up, the degree of convergence keeps decreasing, which indicates degradation of the fitting degree.
- Increasing the epoch number can relieve the over-fitting caused by k.

D. Data Augmentation

- Entity Generation: Even in the same category, the discrete entities, such as actor names, movie names, have very various forms, resulting in data sparseness that is believed to be a potential reason for over-fitting. Our models are supposed to learn the abstract representations of the entity in the specific context, rather than to remember the entity itself. Therefore, we generalized the entities of same category with a unique string, such as "subject_person".
- Knowledge Matching: Assumption of this paper is that knowledge is the key to build human-like chatbot. Using short text matching technology, the specific response relies on the knowledge, which also help to change subject smoothly. So before inputting all the text into encoder, we pre-process knowledge resource, hope to improve the knowledge management module efficiency.

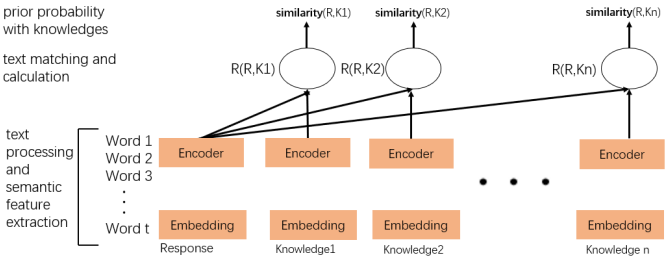

Fig. 4. Knowledge Matching Structure

- Dialogue Extraction: For one sample, if we use all the history dialogue turns as one batch, it means $x_0, x_1, x_2, ..., x_{t-1}$ are inputs, the last response x_t is our output. However, if we define two dialogue turns as a input, the last response is the output. we can extract several samples using the sliding windows.

IV. Experiments

1) Dataset: This paper uses the Chinese DuConv conversation dataset, which includes 30k dialogues, about

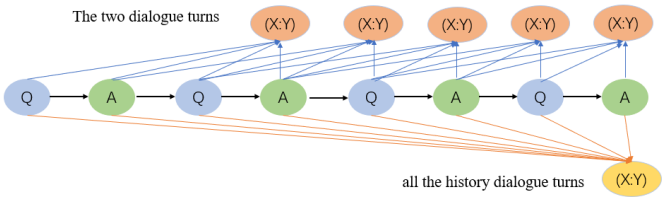

Fig. 5. Dialogue Extraction among conversation turns

120k dialogue [20]. The background knowledge provided in the dataset is collected from the domain of movies and stars, including information such as box offices, directors, reviews and etc., organized as triples {entity, property, value}. The topics given in the dialogue are entities, i.e., movies or stars. The data set includes 30k sessions, about 120k dialogue turns, of which 100k are training set, 10k are development set and 10k are test set. Each conversation is generated by two annotators, one of which plays the agent role and the other plays the user role. The agent was asked to lead the conversation with the given knowledge to achieve the setting goal, and the user just needs to talk without any given information. The agent starts the conversion and talks with the user. The data set includes:

- Part of the Training Data: 400 dialogue turns.
- All Training Data: 100k dialogue turns.
- Development Data: 10k dialogue turns.
- Testing Data: 10k samples.

An example of our data is given below:

```
{
    "goal":
    [
        ["START", "our music", "Ann-Marie"]
        ...
    ],
    "knowledge":
    [
        ["our music", "type", "plot"],
        ["our music", "domain", "movie"]
        ...
    ],
    "conversation":
    [
        "what's the movie name?",
        "our music."
        ...
    ]
}
```

2) Implementation: To evaluate our proposals, we implemented six experiments based on six datasets as D-1, D-2, D-3, D-4, D-5, D-6, which will be subject to different data augment operations. We choose D-1 as our baseline. D-2 is design to test the effect of Loss Regularization. The details can be fond in table I.

This paper uses the dynamic embedding for our model. The dimension of word vector is set to 300 which matches the decoder Gated Recurrent Units (GRU) numbers. And our model is trained by Adam Optimizer with learning rate of 0.01. All the code is implemented by *PaddlePaddle* 1.6.2 on an *NVIDIA TITAN XP* GPU.

TABLE I
DATA SET

Data Augmentation & Loss Regularization	Data Set					
	D-1	D-2	D-3	D-4	D-5	D-6
Entity Generation	√	√		√	√	√
Knowledge Matching				√	√	
Dialogue Extraction					√	√
Loss Regularization		√	√	√	√	√

3) Evaluation: In our experiments, we apply F1, Bilingual Evaluation Understudy (BLEU) [21] and Distinct1/2 as evaluation criteria. F1 is given by the Precision and Recall of test dataset.

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{13}$$

for BLEU-N, the N represents N-grams model of the candidate sentence and the reference sentence, and then calculated by counting the matching number.

$$BLEU = BP * exp(\sum w_n log p_n) \tag{14}$$

$$BP = \begin{cases} 1 & r < c \\ e^{(1-r/c)} & r \geq c \end{cases} \tag{15}$$

In the above formula, r is the length of reference sentence, and c is the length of the candidate sentence. This evaluation method is independent of word order.

In each experiment, we calculate F1, BLEU-1, BLEU-2, DISTINCT-1, DISTINCT-2 scores respectively. The higher F1, BLEU and DISTINCT are, the better the model performs.

TABLE II
EXPERIMENT RESULT

	F1	BLEU1	BLEU2	DISTINCT1	DISTINCT2
D-1	34.97%	0.34%	0.19%	0.03%	0.10%
D-2	36.28%	0.33%	0.18%	0.04%	0.11%
D-3	30.89%	0.30%	0.15%	0.03%	0.11%
D-4	**41.42%**	**0.37%**	**0.22%**	**0.04%**	**0.11%**
D-5	32.89%	0.33%	0.17%	0.01%	0.04%
D-6	27.45%	0.29%	0.12%	0.01%	0.03%

4) Experimental Results and Analysis: The result of our experiments is summarized in Table II, from which we can conclude:

1) Experiment with D-4, which is processed with Entity Generation, Knowledge Matching and Loss Regularization, scores highest among all experiments, compared with baseline D-1. The five criteria improve 7%, 0.03%, 0.03%, 0.01%, 0.1% respectively.
2) Comparing D-2 with D-1, Loss Regularization is proved effective, due to avoiding over-fitting.

3) The comparison of D-2 and D-3 reveals that Entity Generation increases the model's performance. Generalization ability of our model is enhanced.
4) From D-2 and D-4, There is obvious performance increase result from Knowledge Matching which raises the accuracy of the encoder and decoder.
5) The performance of D-5 is worse than D-4, so Dialogue Extraction does not work as we expect. Although increasing the number of samples, it decreases the information of each sample. It is a symbol of under-fitting.

V. Conclusion

Multi-round chatbots fall short in changing the conversation subject smoothly. Knowledge selection mechanism are adopted to address this problem. The innovation of this paper is combination of this mechanism with data augmentations which are proved success in computer version. We propose three Data Augmentation methods including Entity Generation, Knowledge Matching and Dialogue Extraction. Our experiments witness that Entity Generation, Knowledge Matching enhance the model's performance. The first enhances generalization ability of our model, the second increases the accuracy of the encoder and decoder. Meanwhile, Loss Regularization is proved effective, due to avoiding over-fitting. However, Dialogue Extraction is not effective as we expect. Although increasing the number of samples, it decreases the information of each sample which leads to under-fitting.

References

[1] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, "Learning to select knowledge for response generation in dialog systems," *arXiv preprint arXiv:1902.04911*, 2019.

[2] B. Mutlu, T. Kanda, J. Forlizzi, J. Hodgins, and H. Ishiguro, "Conversational gaze mechanisms for humanlike robots," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 2, pp. 1–33, 2012.

[3] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv preprint arXiv:1605.07683*, 2016.

[4] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[5] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[6] Y. Song, R. Yan, C.-T. Li, J.-Y. Nie, M. Zhang, and D. Zhao, "An ensemble of retrieval-based and generation-based human-computer conversation systems." 2018.

[7] A. D' Innocente, F. M. Carlucci, M. Colosi, and B. Caputo, "Bridging between computer and robot vision through data augmentation: a case study on object recognition," in *International Conference on Computer Vision Systems*. Springer, 2017, pp. 384–393.

[8] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention." in *IJCAI*, 2018, pp. 4623–4629.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[10] W. Blacoe and M. Lapata, "A comparison of vector-based representations for semantic composition," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 2012, pp. 546–556.

[11] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[12] J. Tan, X. Wan, and J. Xiao, "From neural sentence summarization to headline generation: A coarse-to-fine approach." in *IJCAI*, 2017, pp. 4109–4115.

[13] P. Li, W. Lam, L. Bing, and Z. Wang, "Deep recurrent generative decoder for abstractive text summarization," *arXiv preprint arXiv:1708.00625*, 2017.

[14] L. Wang, J. Yao, Y. Tao, L. Zhong, W. Liu, and Q. Du, "A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization," *arXiv preprint arXiv:1805.03616*, 2018.

[15] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[16] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[18] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[19] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," *arXiv preprint arXiv:1703.10960*, 2017.

[20] baidu, "Knowledge driven conversation competition," 2019. [Online]. Available: https://aistudio.baidu.com/aistudio/competition/detail/14

[21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu,

"Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.