Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1598–1602
Marseille, 11–16 May 2020

Learning the Human Judgment for the Automatic Evaluation of Chatbot

Shih-Hung Wu*, Sheng-Lun Chien
Department of Computer Science and Information Engineering, Chaoyang University of Technology,
Taichung, Taiwan (R.O.C)
{shwu, s10727614}@cyut.edu.tw
*Contact-author

Abstract

It is hard to evaluate the quality of the generated text by a generative dialogue system. Currently, dialogue evaluation relies on human judges to label the quality of the generated text. It is not a reusable mechanism that can give consistent evaluation for system developers. We believe that it is easier to get consistent results on comparing two generated dialogue by two systems and it is hard to give a consistent quality score on only one system at a time. In this paper, we propose a machine learning approach to reduce the effort of human evaluation by learning the human judgment on comparing two dialogue systems. Training from the human labeling result, the evaluation model learns which generative models is better in each dialog context. Thus, it can be used for system developers to compare the fine-tuned models over and over again without the human labor. In our experiment we find the agreement between the learned model and human judge is 70%. The experiment is conducted on comparing two attention based GRU-RNN generative models.

Keywords: Chatbot, Dialogue generation, learning for evaluation

1. Introduction

Assessment of open domain dialogue systems usually rely on human, and it is hard to have a consistent evaluation. Due to the lack of automatic metrics, there is very limited reproducibility of dialogue system evaluation (Fokkens et al., 2013). At the same time, human evaluation methodologies are also very diverse. Papers report novel generation methods for dialogue systems, but pay little attention on datasets and the evaluation process. Traditional automatic evaluation measures on NLP applications, such as BLEU for machine translation (Papineni et al., 2002), requires human references as ground truth. Recent research on natural language generation makes Chatbot more interesting, however, it is hard to evaluate the quality of the generated dialogue since it is hard to provide human references. Although there are many research on dialogue evaluations (Shawar and Atwell, 2007), most automatic measure metrics cannot reflect the quality of a Chatbot. Currently, the Chatbot evaluation can only rely on human judge, no matter it is for research such as a shared task or for application development (Chen et al., 2019).

There are two main reasons on why it is hard to do the automatic evaluation. First at all, it is hard to evaluate natural language of most natural language processing tasks, such as machine translation and summarization. It required predefined reference for the evaluation and golden reference usually do not exist. The second reason is more complicate, since any known automatic evaluation tool can be incorporated into the generation process; the performance will be no different among these systems that incorporated a known automatic evaluation tool. For example, there are some End-to-End NLG systems that can learn sentence planning and surface realization from non-aligned data (Sedoc et al., 2019). These systems are based on parallel datasets, without the need of human references. However, if an evaluation is fully automatic, then it can be incorporated into a generation system. Thus, it can always generate sentences with higher evaluation value. The evaluation is also based on human judgment. However, it is very hard to develop new models without knowing the feedback during the system developing process. Therefore, we have to compare our own models by ourselves. We used a model A versus model B voting system (A/B test) to acquire the comparison between two models. We show two generated replies for each input post to human judges and collect the human voting results of the pairwise comparisons on which reply is preferred. It is possible to reduce the effort of human evaluation for such kind of comparison by learning a model that can make the same judgment. In this paper, we test the evaluation method on two generative models. The two generative models are based on an attention model GRU-RNN network, and the judgment model is based on BERT. In the following sections, we will introduce the generative models, judgment model, experiments and results, and give conclusion.

In this paper, we will report our experiments on self-evaluation when developing our system in a stared task. In NTCIR-14 STC-3 shared task, there is a sub-task called Chinese Emotional Conversation Generation (CECG) (Zhang and Huang, 2019), which emphasize on dialogue generation with emotion. Table shows three examples, where the post is the input sentence, and the generated response must in the specified emotion category. The goal of the emotion generation subtask CECG is to investigate how well a Chatbot can express feelings by generating a text response to an input. The task is defined as follows: given a post and pre-specified five emotion classes, a system has to generate five responses that are appropriate to both the topic and emotions (Zhou et al., 2018). Table 1 shows some examples. The evaluation is also based on human judgment. However, it is very hard to develop new models without knowing the feedback during the system developing process. Therefore, we have to compare our own models by ourselves.

We used a model A versus model B voting system (A/B test) to acquire the comparison between two models. We show two generated replies for each input post to human judges and collect the human voting results of the pairwise comparisons on which reply is preferred. It is possible to reduce the effort of human evaluation for such kind of

comparison by learning a model that can make the same judgment. In this paper, we test the evaluation method on two generative models. The two generative models are based on an attention model GRU-RNN network, and the judgment model is based on BERT.

The research question of this paper can be defined as: given a limited judge result of the comparison between two system, can the ML model learn and predict the judgment with high accuracy? In the following sections, we will introduce the generative models, judgment model, experiments and results, and give conclusion.

Post	Emotion Category	Reply
爱狗还会做饭的男人，最帅了！ The man who cooks and loves dogs is very handsome!	喜欢 Like	会做饭的男人是很帅的啊。 The man who cooks is handsome.
本来想学一把沧桑，结果令我更忧伤。 I wanted to learn the vicissitudes of life, but I became sadder.	悲伤 Sadness	这是一个悲伤的故事。 It a sad story.
今天一整天都在海边训练。虽然很累，但还是很开心的！ I have been training at the seaside all day. Though very tired, I still very happy!	快乐 Happiness	是的呢，开心！ Yeah, happy!

Table 1: Dialogue example from CECG sub-task

2. Generation Method for Emotion Conversation

This section we will introduction our dialogue generation system.

2.1 Training Set for Emotion Conversation Generation

In NTCIR-14 STC-3 CECG sub-task, the organizers provided a dialogue dataset. The dataset is constructed from Weibo posts and replies (called "comments" in the context). Above one million Weibo post-response pairs is provided for training. The test dataset consists of about 5000 posts while 200 of the posts are manually assessed, and for each post five replies for five different emotion are generated.

The dataset includes emotion labels of each post and response. However, these labels are for reference only, and they are obtained by a simple classifier based on a bidirectional LSTM model. The classifier was trained on the data from the NLPCC Emotion Classification Challenge and the accuracy for six-way classification is about 64%. Thus, the emotion label of these data is no accurate.

2.2 Generation Model

Figure 1 shows a text generation system based on a bi-directional GRU-RNN network including an attention model, which was proposed for neural machine translation (Luong, 2015)(Bahdanau et al., 2015). We train the model to generate a response in specified emotion for a given input sentence. To generate response in specific emotion, our generator system GRU-RNN network includes an emotion tag [EM] as part of the input. The emotion tag will be: 0: Other, 1: Like, 2: Sadness, 3: Disgust, 4: Anger, 5: Happiness.

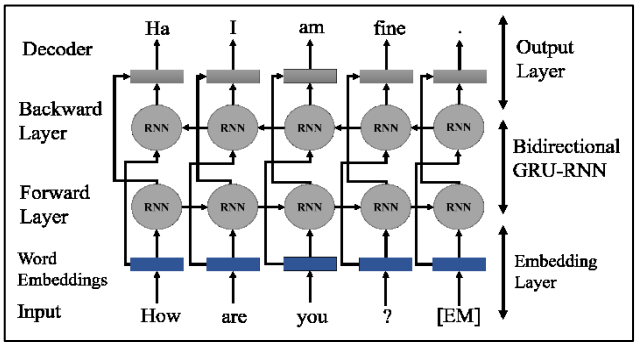


Figure 1: Our bidirectional GRU-RNN text generation model

The goal of emotional sentence generator in NTCIR14-STC3 CECG is given a Chinese post $X = (x_1, x_2, \cdots, x_n)$ and an emotion class, the system will generate a response $Y = (y_1, y_2, \cdots, y_m)$ for each input post in that is coherent with the emotion category. The text generation training data set is constructed from Weibo posts and replies. More than one million Weibo post-response pairs are include in our training data set (Huang et al., 2017). Our system is built without outside knowledge other than the training set. We do not use pre-train embedding and word embedding is simple done with the Pytorch torch.nn.Embedding() method[1]. Two dialogue examples are shown in Table 2.

Post 1	為什麼 為什麼 為什麼 你們 都 不 陪 我 看電影 !! Why why why you guys are not accompanying me to the movies!!
Reply 1	因為 你 不 喜歡 嘛 Because you don't like it.
Post 2	老骨頭 一把...實在 打 不 動 球 了... I am old bones... I really can't play the ball anymore ...
Reply 2	哈哈 你 真 幽默 Haha you are so funny.

Table 2: Examples of the input post and the generated reply by our system

By setting different hyperparameters, we can create similar model of this kind of text generation model that can generate different results. For example:

Model A: Limit the number of output words to 5

Model B: Limit the number of output words to 50

[1] https://pytorch.org/docs/stable/nn.html

However, it is hard to know which is better. We have to hire human judge to compare the generated results from these two models for each post. It takes human labor and time to collect the voting results on which model is preferred for each input post. Each model generated 1,000 response for the official CECG sub-task, and we collected the human judgment as a kind of training data for our automatic evaluation model to learn. Figure 2 shows how we generate two different responses to the same input post and Figure shows how we collect the human judgment.

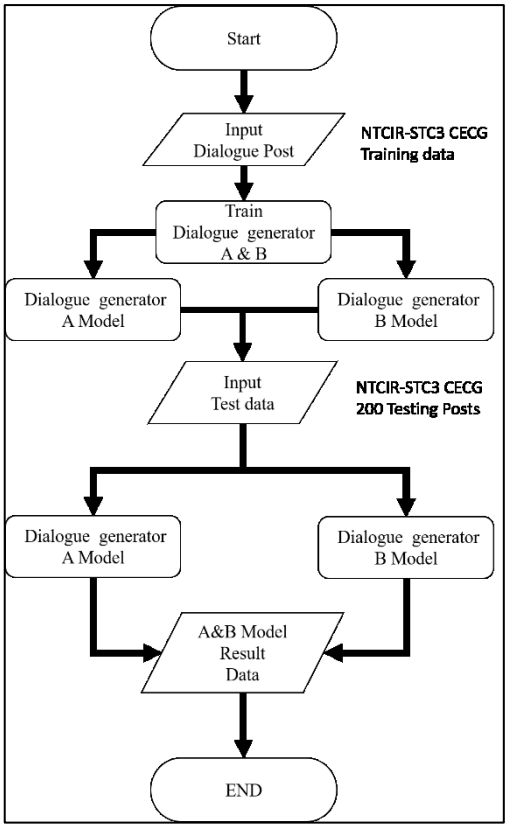


Figure 2: How we generate two different responses to the same input post

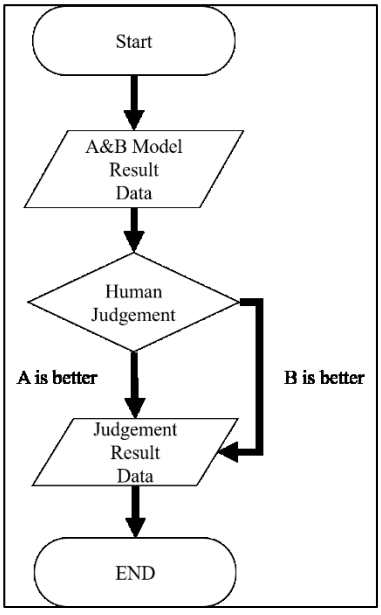


Figure 3: How we collect the human judgment

There are two dialogue generators named model A and model B, and we use each model to generate 1,000 responses. We hired four graduate students as the human judge for the human evaluation. They have to choose which response is better than another between two generated sentences to the same input post.

Post 1	这 作业 做 个 通宵 都 做 不 完 。 谁 来 陪 我 聊聊 免得 我 睡着 了 [泪] This homework can't be done even overnight. Who can chat with me to avoid me fall in sleep [tears]	
	Model A	Model B
Reply 1	辛苦 了 Hard work	**你 是 不是 该 做 作业 了** **Should you do your homework**
Post 2	刚 吃 了 早 午餐 , 开始 劳作 啦 [兔子] Just had brunch and time to work [rabbit]	
	Model A	Model B
Reply 2	**祝 你 好运** **Wish you luck**	我 喜欢 吃 甜 的 馄饨 I like to eat sweet dumplings

Table 3: Examples of the test result from model A and model B model. In the first case, for Post 1, user prefers reply 1 from model B, and in the second case, for Post 2, user prefers reply 2 from model A.

3. Automatic Evaluation by Learning the Human Evaluation with BERT

In this section we present our main contribution. We try to build a model that can judge which model is preferred based on the voting of human. If the machine learns how to evaluate the text like human, it can do the automatic evaluation. We created an automatic evaluation model based on the BERT (Devlin et al., 2018) pre-trained language model and a simple linear classification that can classify which generated sentence is preferred as shown in Figure 4.

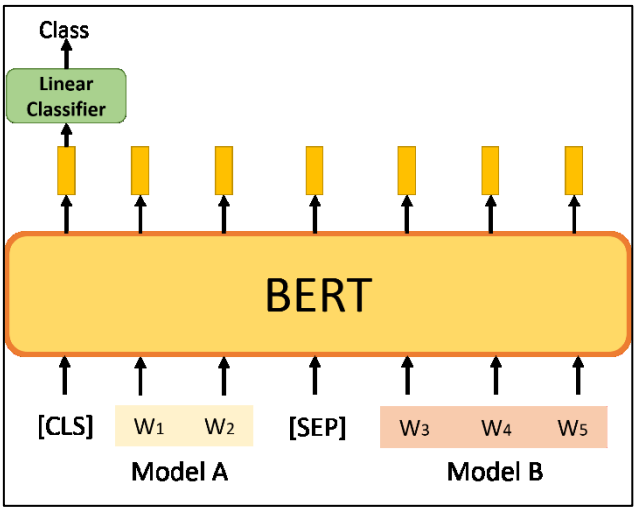


Figure 4: Our automatic evaluation model based on BERT

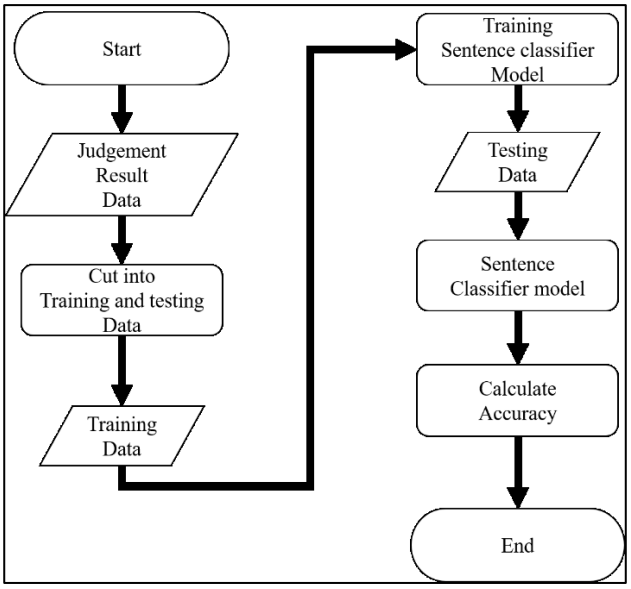
In our experiments, the input sequence of Model A is the post plus the generated response by Model A. The input sequence of Model B is the post plus the generated response by Model B. The target class is the human preference on Model A or Model B, given the post

4. Experiments

4.1 Experimental Setting

This section we will explain how the experiment goes. We design two way to test how many training examples are necessary to get a stable model, and see how well it can work. The model is built on various training set size from 100 to 900 and use the rest data set as test data. In the first setting, we used the data in the original sequence. In the second setting, we shuffled the data before the experiments. Our system process flow chart is as follow Figure 5.

Figure 5: How we train and test the automatic evaluation model



4.2 Experiment Results

The experiment results are shown in Figure 6 and Figure 7. As we can see in Figure 6 and Figure 7 that the model can predict about 70% of the human judgment with only limited train set. In the original data setting, the performance increasing unstably, we believed this is due to the natural instability of human preference. The human judge cannot give consistent judgment. In the shuffled data setting, the performance increasing stably, since data shuffling eliminate the instability. In Figure 7, when the training set size at 800, the system perform best using all the other 200 sentences as the test set. The accuracy is very high, considering a random based line is 50% and with such small training set.

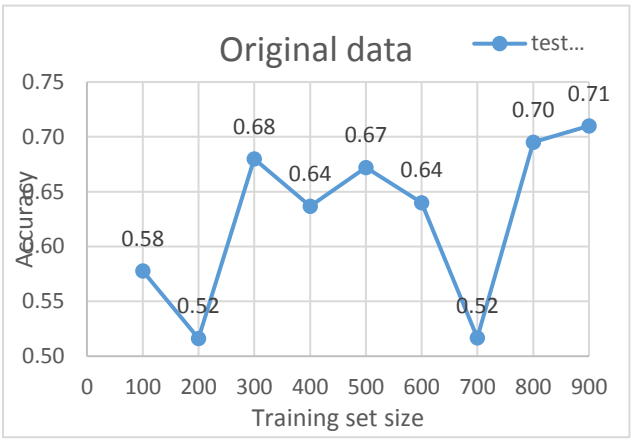


Figure 6: The test result of the automatic evaluation model

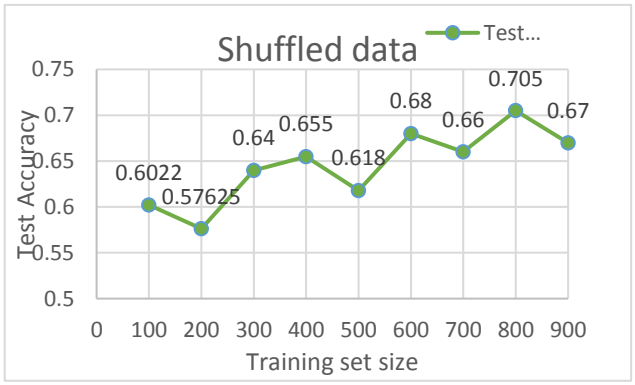


Figure 7: The test result of the automatic evaluation model

5. Conclusion

In this paper we proposed an evaluation method that can reduce the cost of human labor on dialogue evaluation. The main contribution is decreasing human labor and create a stable automatic evaluation tool. We created an automatic evaluation model based on the BERT (Devlin et al., 2018) pre-trained language model and a simple linear classification that can classify which generated sentence is preferred. The model can predict about 70% of the human judgment with only limited train set.

Traditionally, text are represented by the meaning of the content words, from bag-of-words model to word embedding. The quality of a sentence lies on other issues, such as vocabulary, spelling, grammar usage, writing styles, and insight on the topic and how to communicate ideas to the readers. It is hard to compare two sentences with only via content. Two responses might using the same words, but with some subtle difference, such that one might be moving while another is boring. In this paper, our experimental result shows that non-content quality can be modelled to some degree.

6. Acknowledgements

This research is partially sponsored by Chaoyang University of Technology (CYUT) and Higher Education Sprout Project, Ministry of Education, Taiwan, under the project name: "The R&D and the cultivation of talent for Health-Enhancement Products".

7. Bibliographical References

Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473.

Chen, X., Mi, J., Jia, M., Han, Y., Zhou, M., Wu, T. and Guan, D., (2019), Chat with Smart Conversational Agents: How to Evaluate Chat Experience in Smart Home. 1-6. 10.1145/3338286.3344408.

Devlin, j., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2 [cs.CL]

Fokkens, A., Erp, M., Postma, M., Pedersen, T., Vossen, P. and Freire, N. (2013). Offspring from reproduction problems: What replication failure teaches us. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 1691–1701.

Luong, M-T., Pham, H. and Manning, C.D. (2015). Effective approaches to attention based neural machine translation. arXiv preprint arXiv:1508.04025.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation, in Proceedings of ACL, pp 311-318.

Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L., Callison-Burch, C. (2019). ChatEval: A Tool for Chatbot Evaluation, in proceedings of Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), June, Minneapolis, Minnesota, pp.60–65.

Shawar, B. and Atwell, E. (2007). Different measurements metrics to evaluate a chatbot system, Proc. of the 2nd Workshop on TextGraphs:Graph-Based Algorithms for Natural Language Processing. 2007.89-96.https://doi.org/10.3115/1556328.1556341.

Zhang, Y. and Huang, M. (2019). Overview of the NTCIR-14 Short Text Generation Subtask: Emotion Generation Challenge, in Proceedings of NTCIR-14, Tokyo, Japan, June 10-13.

Zhou, H., Huang, M., Zhu, X., and Liu, B. (2018), Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. AAAI 2018, New Orleans, Louisiana, USA. Previous version at arXiv:1704.01074.

8. Language Resource References

Huang, M., Ye, Z., Zhou, H. (2017). Overview of the Emotional Conversation Generation Challenge Task at NLPCC. NLPCC 2017.