Evaluation Methods of Emotional Expression in Short Text Conversation

Shih-Hung Wu†, Wen-Feng Shih, Sheng-Lun Chien
Department of Computer Science and Information Engineering
Chaoyang University of Technology
Taichung, Taiwan (R.O.C)
shwu@cyut.edu.tw, wu0fu491@gmail.com, s10727614@cyut.edu.tw

ABSTRACT

With the advance of the study on automatically generated conversation, the research on evaluation is also getting important. How to evaluate the quality of the emotional conversation text is our research goal.

The two major evaluation methods have their own drawbacks. The automatic evaluation methods can judge the dialogue system quickly; however, there is no commonly accepted metrics currently. On the other hand, human judgments suffer from inconsistency; the inter-annotator agreement is unstable. In this paper, we conduct a study on how to make the human judgment more stable by analyzing the mutual agreement between different human judges, and discuss how to systematically design evaluation questions.

We discuss how to improve the evaluation rules in STC-2 task in NTCIR, which originally are not designed for emotional conversation. We design a process to find out stable factors with catharsis emotional aspects to improve the evaluation rules for emotional dialogue evaluation. The dialogue data with catharsis types are gathered from our STC-2 system. Evaluation questionnaire with different aspects is verified on whether it can achieve consistency or not. By analyzing the questionnaire survey result, we find the aspects that can achieve higher consistency.

CCS CONCEPTS

•Information retrieval •Evaluation of retrieval results

KEYWORDS

Short text conversation, Natural language generation, Conversation generation evaluation, Catharsis dialogue.

1 Introduction

Social Chatbot that can express emotions is getting popular recently; motional social Chatbot can be used in many situations such as elderly care or babysitting. Our motivation began with the NTCIR short text dialogue (short Text Conversation, STC); we participated in STC, and STC2 [1][2]. In the first STC task, Sina Weibo was used as the data set for Chinese, and we considered STC as an information retrieval (IR) task. STC-2 added a generation project, and the system had to generate ten comments to each post, so it is more difficult than IR. The dialogue in STC is very similar to the way a Social Chatbot needs to communicate with people. No matter we use IR technology to retrieve the appropriate sentences or use generation method to generate sentences, evaluation of the sentences rely on human.

1.1 Automatic Evaluation Is Not Ready

Currently, most of the researches on the Chatbot dialogue system are focused on the development of the system, i.e. technology and conditions that make better sentences and more appropriate responses to the topic. An effective evaluation method that can measure the response is not ready. At present, the evaluation method is divided into two main categories, automatic evaluation and manual evaluation.

The purpose of automatic evaluation is to automate the evaluation of the appropriateness of the returned or generated sentences. Although it is possible to quickly evaluate the dialogue system and give a score, the automatic evaluation often results in different scores from human evaluation. It is difficult to use automated assessments in the evaluation dialog system.

For example, BLEU [3] is a way to evaluate the quality of machine-translated text based on n-gram matching. It is one of the popular automated evaluation matrices. In the case of machine translation, BLEU is handy, but in a conversation system, a reply sentence may not overlap a lot to the reference text and results in a low BLEU score. Therefore, evaluation based on overlap cannot be used to evaluate dialogue system.

In Liu et al [4], the difference between automated and manual assessments is studied. As shown in table 1, where text-overlapping automated assessments give poor scores in correlation coefficient to the manual assessments in two dialogue datasets. This study

† Contact Author

9th International Workshop on Evaluating Information Access (EVIA 2019), co-located with NTCIR-14, 10 June 2019, Tokyo, Japan.

suggested that it is better to use manual evaluation to evaluate the *Chatbots'* dialogue system.

Table 1: Automatic evaluation methods get low correlation score [4]

	Twitter				Ubuntu			
Metric	**Spearman**	p-value	**Pearson**	p-value	**Spearman**	p-value	**Pearson**	p-value
Greedy	0.2119	0.034	0.1994	0.047	0.05276	0.6	0.02049	0.84
Average	0.2259	0.024	0.1971	0.049	-0.1387	0.17	-0.1631	0.10
Extrema	0.2103	0.036	0.1842	0.067	0.09243	0.36	-0.002903	0.98
METEOR	0.1887	0.06	0.1927	0.055	0.06314	0.53	0.1419	0.16
BLEU-1	0.1665	0.098	0.1288	0.2	-0.02552	0.8	0.01929	0.85
BLEU-2	0.3576	< 0.01	0.3874	< 0.01	0.03819	0.71	0.0586	0.56
BLEU-3	0.3423	< 0.01	0.1443	0.15	0.0878	0.38	0.1116	0.27
BLEU-4	0.3417	< 0.01	0.1392	0.17	0.1218	0.23	0.1132	0.26
ROUGE	0.1235	0.22	0.09714	0.34	0.05405	0.5933	0.06401	0.53
Human	0.9476	< 0.01	1.0	0.0	0.9550	< 0.01	1.0	0.0

1.2 Human Evaluation in STC-2

However, the consistency of manual assessments is a problem. For manual evaluation, how to design the assessment questions is important. We first study the STC-2's assessment methodology, which was not designed for specific task or emphasized on empathy. The evaluation rules for STC-2 are shown in Figure 1 and have four conditions, namely fluent (fluent, smooth), Coherent (whether it is consistent with post content), Self-sufficient (whether it can become a separate post for itself) and Substantial (whether it can provide new information for the post). We can view it as asking a human assessor four questions on each one reply comment. In this rule, the returned sentences are divided into 3 levels, L0, L1, and L2. If the reply comment is not fluent, smooth, or consistent with the post content, it will be given the L0 label. If a sentence that satisfies the preceding conditions but not conform to a separate sentence or provides new information to post, it will be the L1. Only when all of the above are met will it be assigned L2. On the basis of the evaluation rules, STC organizers give a score, L0~L2 to each comment. A system is evaluated by three evaluation metrics calculating all commends submitted by the participant. The first is nG@1 [1]. The second is nERR@10, which has the characteristic of diminishing benefit values and is a popular measure [22]. The third is P+, proposed by SAKAI, Tetsuya [23], similar to ERR.

```
IF (fluent AND coherent)
    IF (self-sufficient AND substantial)
        assign L2
    ELSE
        assign L1
ELSE
    assign L0.
```

Figure 1: STC-2 evaluation rules

Let's observe one example on how it works on one test post and two reply comments in Table *2. In T*able 2, there is a post and two comments. *W*e can find that the scores of these two comments are both L2, but *these two comments are emotionally* different *for us*. Comment 1 can be *a* kind of consolation, while Comment 2 is a little bit mocking. The dialogue with an emotional exchange should be included in the assessment.

The assessment is important to guide the direction of the system development. So we try to improve the STC-2 scoring method, and take the emotion exchange into consideration.

Table 2: Test examples with the same label but different emotions

Post	在这边过得太焦虑，天天加班，周末加班，没完没了，午觉也睡不安心 (Too anxious on this side, overtime every day, weekend overtime, endless, nap also sleep not at ease)	**Labels**
Comment 1	祝你周末不加班，周一加休一天！(Wish you no overtime on weekends, one day off in Monday!)	L2
Comment 2	祝你周末加班，天天加班有动力，哈哈哈哈哈(Wish you overtime on weekends, work overtime every day is motivated, hahaha haha)	L2

1.3 Chatbot systems

Chatbot is a system that uses natural language to interact with users, a kind of human-machine interaction, its development history from 1960 ' s. Joseph Weizenbaum's Eliza [7] is the earliest Chatbot. Its purpose was to imitate the way psychotherapists speak, and made people believe that it is a true human being. The goal of these early Chatbot systems was to pass the Turing test [8]. Chatbot system has many applications; it can imitate people's talk. The interaction with Chatbots is generally text-based; recently some Chatbots also use voice-based communication. Zadrozny, et al. [9], said the best way to promote human-machine interaction for users was to express their interests, desires, and questions through voice, typing, or finger pointing. In the future society, a variety of applications will emerge. Argal, Ashay, et al. [10] proposed that an intelligent travel robot, based on DNN technology, can help users search for inquiries about the place of travel and analyze it to give a suitable response. It is equipped with voice interaction, and it can make the interaction work better.

One STC-2 participant system, SG01 [11], used a number of features to match candidate sentences, and the use of learning ranking algorithm to obtain higher ranking results. The SPLAB team [12] builds on the method, using the encoder-decoder framework to develop the system. However, what the model produce is a short and boring response, in order to solve the problem, the multi-resolution recursive neural network and external memory based sequence generation method is proposed. There are many teams based on the method Sequence to Sequence (Seq2Seq) methods to build Chatbots [13][14]. With excellent results in machine translation and natural language generation, Seq2Seq [15][16] is a rapidly developing generation model in recent years. Long Short-Term Memory (LSTM) is one of Seq2Seq's architectures that solves the problems encountered by deep neural networks (DNNs) in sequence pairs, Because DNNs can only be used in fixed dimensions in input and output. The Seq2Seq model,

which consists of two LSTM. The first LSTM reads the part of the sequence, one word at a time to get a large dimension vector. The second LSTM is to extract the output sequence from the vector. Since the second LSTM is based on the recursive neural network (RNN) language model, it can learn long-time dependent data capabilities with good results. There are also other mechanisms based on the Seq2Seq model to achieve better results. Zhang, Ruqing, et al. [17] used Gaussian Kernel layer to guide the model to generate different responses at different specific levels. By adding new parameters to the Seq2Seq model, the author can make the length of the resulting longer, and the content of the reply can be more vivid and interesting.

2 Background

The research strategy is at first we identify the emotional exchange in the dataset of our system output in the STC-2, and evaluate them with different sets of assessment questions. Then we measure the evaluation results, especially the consistency.

2.1 Consistency Measurements

In order to have a reliable assessment, repetitive measurements are necessary. There are two kinds of reliability, firstly the reliability between several persons (Inter-rater Reliability) and secondly the reliability of one person (Intra-rater Reliability). The former is whether there is a consistent view of the same thing when there are more than two evaluators. The latter is to assess whether the same evaluator is consistent with the results of a repeated assessment of the same sample. The common method to evaluate consistency is the Cohen's kappa[18], which can be used to calculate the consistency of things between two evaluators. This kappa value is divided into non-weighted and weighted versions. When we have more than two evaluators, we can't use Cohen's kappa to calculate consistency. Fleiss' kappa [19] can be used when the evaluators are more than 2. The Pearson correlation coefficient [20] is the linear correlation between two variables X and Y with values between -1~+1. So this coefficient can be used to calculate the correlation between two evaluators. The Spearman's rank correlation coefficient [21] is another correlation between the two variables.

2.2 Our IR System in STC-2

Our system architecture is described in previous works. It is mainly divided into several parts, including how to sorting of data, pre-process, word segmentation, index, retrieval and filtering [5][6]. Table 3 shows one official test post of STC-2 and the 10 reply comments from our system. The total number of STC formal test posts is 100 sentences, each test post with 10 reply comments from our system. Our system retrieves 1000 sentences, which will be used in the questionnaire survey.

Table 3: Examples of the results of the formal run of our system

Post	我眼睛的度数好像又增加了，完全看不见你的智商了 (The degree of my eyes seems to have increased again, and I can't see your IQ at all.)
Comment 1	智商和眼睛度数成正比。(IQ is proportional to the degree of the eye.)
Comment 2	抱歉，眼睛近视，度数也有点高，看不见。感情是甚么？(Sorry, eye myopia, degree is also a bit high, invisible. What are the feelings?)
Comment 3~10	…

3 Methodology

In this section, we will introduce our proposed methodology on how to design a set of assessment questions.

3.1 Questionnaire Design Process

Figure 3 is our questionnaire design process, in which we observe the original evaluation rules and analyze the components involved in the post-comment pairs from the corpus, and then combine these newly design-oriented assessment questions with the original evaluation rules to make a new assessment questionnaire. After we conduct the questionnaire survey, we use the Cohen's Kappa value to assess the degree of consistency of the assessment questions we have designed. Assessment question with low kappa values should be removed. This process will not only find effective assessment questions, but also make the questionnaire easier for human judges.

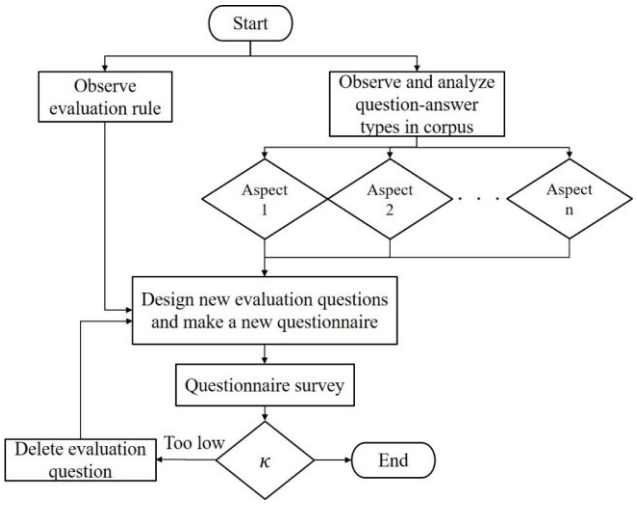

Figure 2: Questionnaire design process

3.2 Topic Survey

In order to know what type of emotion exchange to analyze in the STC-2 data, we listed several types of conversation and surveyed college students among what type of topic they were willing to chat with a chatbot, as shown in Figure 3.

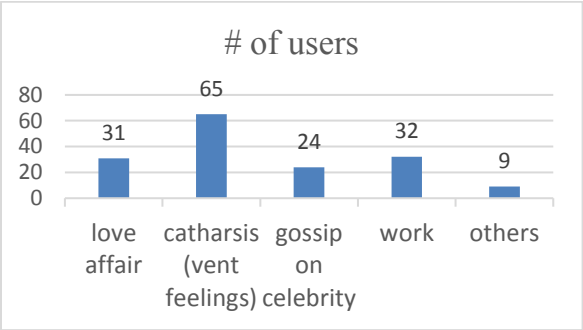

Figure 3: Topics that users preferred to chat with a Chatbot

In figure 3, we can find that most students are most willing to chat with the Chatbot on catharsis type. The conversation is usually used to vent and pour their own negative energy to achieve the effect of compression release, which is an important part of the chat, and the reply of the sentence will directly affect the mood of the receiver. So we concentrate on the assessment question for the catharsis type. One example of conversation in this type is shown in table 4.

Table 4: Conversation example in catharsis type post

Post	假期沒人約会的心情. . . 你们不会懂。。 (The mood of no one dating during the holidays ... You don't understand.)
Comment	会好的。慢慢调节。心情不好的时候总会过去。 (It's going to be okay. Adjust slowly. When you are in a bad mood, you always pass.)

We manually screen the reply sentence as a new corpus, and observed the corpus. We find that the reply comments are divided into three kinds of aspects: Sympathy, Mocking and Consolation. Therefore, we will include the three types of aspects into the assessment questions. Table 5 provides examples.

Table 5: Different aspects replies on catharsis type post

Post	Comment	Aspects
为什么我的眼里常含泪水，因为没人约我出去吃饭。 (Why do I always have tears in my eyes, because no one asked me out for dinner.)	为什么眼里含满泪水？！ (Why is the eyes full of tears?!)	Sympathy
我不是胖，我是骨架大…… (I'm not fat, I'm a big skeleton ...)	是啊，一吃就胖得伤不起…… (Yes, it's too fat to hurt when you eat it ...)	Mocking
假期沒人約会的心情. . . 你们不会懂。。(The mood of no one dating during the holidays ... You don't understand.)	会好的。慢慢调节。心情不好的时候总会过去。(It's going to be okay. Adjust slowly. When you are in a bad mood, you always pass.)	Consolation

3.3 Assessment Questions

Together with STC's assessment questions, we create new questionnaires. Table 6 is our assessment question, a total of 10 questions. We design six assessment questions, the 4th, 5th and 6th question are corresponding to the three aspects. In table 9, A is a post, and B means a reply comment.

Table 6: Assessment questions

	Six our assessment questions	Four STC-2 assessment questions
1	Is this a casual reply?	Whether the response is fluent?
2	Is this an earnest reply?	Is B's response consistent for A?
3	Is the reply deviating from the topic?	Can B's response be regarded as a single message?
4	Is it a sympathy reply?	Does B's response to whether A provide new information?
5	Is it a consolation reply?	
6	Is it a mocking reply?	

3.4 Inter-evaluator Agreement

In this section, we will describe the evaluation method used to calculate the kappa value. There are two main types, namely Cohen's Kappa and Fleiss' Kappa. The difference between the two is the number of evaluators; Cohen's Kappa is to assess the consistency between two evaluators, while Fleiss' Kappa can assess the consistency among multiple evaluators. The interpretation of kappa value is shown in Table 7 [25].

Table 7: The interpretation of kappa value [25]

κ	Agreement Level
0-.20	None
.21-.39	Minimal
.40-.59	Weak
.60-.79	Moderate
.80-.90	Strong
Above .90	Almost Perfect

4 Experiments and Results

We designed two experiments to test the inter-annotator agreement among different assessment questions.

4.1 Experiment One

In this experiment, we use the assessment questions that we defined in section 3 and the STC-2 assessment questions as a new questionnaire, and conduct a questionnaire survey. The purpose of the experiment is to observe whether the additional assessment problems would make it easier to achieve consistency. The subjects of our survey are college students, and the number of questions is 12, of which 6 are background surveys. Our total measured subjects are 83, and the effective questionnaires are 78. Table 8 shows the Kappa of each question in the questionnaire.

Table 8: Fleiss' kappa in experiment 1

	Our assessment questions	STC-2 assessment questions
1	0.246	0.077
2	0.246	0.133
3	0.122	0.025
4	0.245	0.046
5	0.368	
6	0.277	
average	0.250	0.070

From the results, kappa values of our questions are higher than Kappa values of the STC-2 question. As in section 3.3 mentioned, the higher the number of evaluators, the less likely it will get a high Kappa value. The number of evaluators in experimental one was 78, so we were able to get 0.250 of the large number of evaluators Kappa value. In addition, we randomly sampled two copies from 78 questionnaires to calculate Cohen's kappa and to see if there were a high degree of consistency. Table 9, Table 10 are two inter-evaluator matrix, and table 11 is the result of the calculation of Cohen's Kappa.

Table 9: Inter-evaluator agreement matrix (randomly choose two evaluators) of our assessment questions

		evaluator 41	
		Yes	No
evaluator 40	Yes	11	6
	No	5	14

Table 10: Inter- evaluator agreement matrix (randomly choose two evaluators) of STC assessment questions

		evaluator 41	
		Yes	No
evaluator 40	Yes	1	7
	No	6	10

Table 11. Cohen's kappa

	Our assessment questions	STC-2 assessment questions
κ	0.385	-0.258

As in the results of Cohen's Kappa, we can find that the kappa value of the assessment questions we designed is also higher than the assessment of the STC-2 questions. It can be learned from the experiment that the assessment questions of our design can achieve a high degree of consistency.

4.2 Experiment Two

In the second experiment, we use our formal run submitted data as post-comment pairs, and evaluate with the new questionnaire. The ten questions are the same as in the first experiment. The reference of STC indicate that it requires at least three evaluators to evaluate, so we invited 3 evaluators to evaluate the new questionnaire.

4.2.1 STC-2 official evaluation dataset. Table 12 is the data statistics of the STC-2 dataset [2]. The data comes from SINA Weibo's micro-weblog website. During the test phase, the STC-2 organizers give test data for a total of 100 test post and ten comments are required to reply each post.

4.2.2 Experiment 2 result. Although we design the assessment questions for the catharsis type only. We test them on the formal run test post from the STC. The purpose of experiment 2 is to observe whether the assessment questions we have designed is more likely to be consistent when there are various types in the data. Our 3 evaluators are graduated students major in computer science. The number of questionnaire for each evaluator consists 1000 questions. Of the 3 evaluators, we calculated the Cohen's Kappa value for each pair of evaluators, table 13, table 14 was the result of our assessment of the problem with STC in evaluators 1 and 2, and table 15 shows Cohen's Kappa.

Table 12. STC-2 data set statistics [2]

Repository	No. of posts	219,174
	No. of comments	4,305,706
	No. of original pairs	4,433,949
Labeled Data	No. of posts	769
	No. of comments	11,535
	No. of labeled pairs	11,535
Test Data	No. of query posts	100

Table 13. Inter- evaluator agreement matrix (two annotators) of our assessment questions

		evaluator 2	
		Yes	No
evaluator 1	Yes	983	706
	No	695	3616

Table 14. Inter- evaluator agreement matrix (two annotators) of STC assessment questions

		evaluator 2	
		Yes	No
evaluator 1	Yes	619	678
	No	491	2212

Table 15. Cohen's kappa

	Our assessment questions	STC-2 assessment questions
κ between evaluator 1 and evaluator 2	0.421	0.307
κ between evaluator 3 and evaluator 2	0.444	0.336
κ between evaluator 1 and evaluator 3	0.402	0.249

The consistency between Evaluator 1 and 3 can be found in table 26 without the height of the above two groups. However, our assessment problems are also more consistent than those of STC. In addition, we calculate Fleiss' kappa for these 3 evaluators, as shown in table 16.

Table 16. Fleiss' kappa in experiment 2

	Our assessment questions	STC-2 assessment questions
1	0.076	0.146
2	0.097	0.357
3	0.379	0.170
4	0.033	0.058
5	0.121	
6	0.101	
average	0.134	0.182

From Table 16, we can find that the Fleiss' kappa value of our e assessment questions is lower than that of STC-2, and according to this case, we will analyze the distribution of the questionnaire again, and find out why the Cohen's kappa value is high instead of Fleiss' kappa value is low. Table 17 provides the distribution of the same views among the 3 evaluators in the experiment two. From this table, it can be found that the main same distribution is on 3-oriented assessment issues, and the more consistent the Cohen's Kappa value will be because of the higher distribution of the same opinion. However, we analyzed the Fleiss' kappa value and found that when the opinion among the evaluators was almost one side, the Fleiss' kappa value would get 33 a lower value, but the result would be known to the results of Table 16 and table 17. However, in this distribution can be found in our design evaluation problems, the first three assessment questions get less consistency, so we can see that these three questions are relatively bad assessment questions.

Table 17. Numbers of agreement among three evaluators in Experiment 2 (out of 1000)

	Our assessment questions	STC-2 assessment questions
1	518	479
2	477	613
3	467	633
4	909	671
5	958	
6	868	

5 Conclusion and Future Work

We designed a new process to design questionnaire for human assessment on emotional dialogue evaluation. We use Cohen's kappa and Fleiss' kappa to calculate human consistency and determine whether a question is good or not. The experimental results show that the evaluation question of our design can be more consistent than that of STC-2. From the survey results, we found that our assessment questions obtain a high degree of consistency. Our assessment questions are involved in several important emotional aspects in addition to the STC-2 evaluation rules which focused on general dialogue quality.

Just as what we have done on human evaluation, emotion-involved metrics should be explored more. In the future, we need to find out more about the different aspects, especially the emotions about sympathy. Emotions are very important factors in chatting, and emotions can directly affect each other's feelings. Deeper emotions are relative to the human psychological factors, and deeper emotional questions should be added to the evaluation problem when evaluating the conversation system.

Automatic assessment metrics that can replace human evaluation still need to be studied more. Many metrics have been proposed, such as quality, novelty and divergence [26], where quality means the perplexity of a sentence according to a language model, and novelty measures the portion of the generated sentence being a reproduction of an old sentence. The divergence metric is to measure the difference among generated sentences. Stent et al. showed that automatic evaluation metrics can partially measure adequacy (similarity in meaning), but are not good measures of fluency (syntactic correctness) [27]. In STC task, there are more needs on fluency and emotion. Most automatic evaluation metrics are not suitable for the task. Novikova et al. showed that that state-of-the-art automatic evaluation metrics do not sufficiently reflect human ratings, which means human evaluations is necessary [28].

ACKNOWLEDGEMENTS

This study is supported by the Ministry of Science and Technology under the grant numbers MOST106-2221-E-324-021-MY2.

REFERENCES

[1] Lifeng Shang, Tetsuya Sakai, Zhengdong Lu, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao. "Overview of the NTCIR-12 short text conversation task." *In Proceedings of NTCIR-12*, 2016, pp. 473-484.

[2] Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, Masako Nomoto. "Overview of the NTCIR-13 Short Text Conversation Task." *In Proceedings of NTCIR-13*, 2017, pp. 194-210.

[3] Papineni, Kishore, et al. "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. p. 311-318.

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

[5] Shih-Hung Wu, Wen-Feng Shih, Liang-Pu Chen and PingChe Yang, "CYUT Short Text Conversation System for NTCIR-12 STC." *In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, June 7-10, 2016, Tokyo Japan, pp.541-546.

[6] Shih-Hung Wu, Wen-Feng Shih, Che-Cheng Yu, Liang-Pu Chen and PingChe Yang, "CYUT-III Short Text Conversation System at NTCIR-13 STC-2 Task. " *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, December 5-8, 2017, Tokyo Japan, pp.289-294.

[7] Weizenbaum, Joseph. "ELIZA—a computer program for the study of natural language communication between man and machine." *Communications of the ACM,* 1966, vol. 9, no. 1 pp. 36-45.

[8] Turing, Alan M. Computing machinery and intelligence. In: *Parsing the Turing Test*. Springer, Dordrecht, 2009. p. 23-65.

[9] Zadrozny, W., Budzikowska, M., Chai, J., Kambhatla, N., Levesque, S., & Nicolov, N. "Natural language dialogue for personalized interaction." *Communications of the ACM,* 2000, vol. 43, no. 8, pp. 116-120.

[10] Argal, Ashay, et al. Intelligent travel chatbot for predictive recommendation in echo platform. In: *Computing and Communication Workshop and Conference (CCWC), 2018 IEEE 8th Annual*. IEEE, 2018. p. 176-183.

[11] Zhao, H., Du, Y., Li, H., Qian, Q., Zhou, H., Huang, M., & Xu, J. "SG01 at the NTCIR-13 STC-2 Task. " *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, December 5-8, 2017, Tokyo Japan, pp.313-316.

[12] Liu, X., Wu, X., Chen, R., Zhao, Z., Lin, H., & Yu, K. "splab at the NTCIR-13 STC-2 Task." *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, December 5-8, 2017, Tokyo Japan, pp.282-288.

[13] Yihan, L., Shanshan, J., Lei, D., Yixuan, T., & Bin, D. "SRCB at the NTCIR-13 STC-2 Task." *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, December 5-8, 2017, Tokyo Japan, pp.237-240.

[14] Nakatani, H., Nishiumi, S., Maeda, T., & Araki, M. "KIT Dialogue System for NTCIR-13 STC Japanese Subtask." *In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, December 5-8, 2017, Tokyo Japan, pp.257-264.

[15] Sutskever, Ilya; Vinyals, Oriol; Le, Quoc V. "Sequence to sequence learning with neural networks." In: *Advances in neural information processing systems*. 2014. p. 3104-3112.

[16] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[17] Zhang, R., Guo, J., Fan, Y., Lan, Y., Xu, J., & Cheng, X. "Learning to Control the Specificity in Neural Response Generation." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018. p. 1108-1117.

[18] Fleiss, Joseph L., and Jacob Cohen. "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability." *Educational and psychological measurement,* 1973, vol. 33, no. 3 pp. 613-619.

[19] Fleiss, Joseph L. "Measuring nominal scale agreement among many raters." *Psychological bulletin,* 1971, vol. 76, no. 5 pp. 378-382.

[20] Benesty, J., Chen, J., Huang, Y., & Cohen, I. "Pearson correlation coefficient." *Noise reduction in speech processing*. Springer, Berlin, Heidelberg, 2009. 1-4.

[21] Myers, Jerome L.; Well, Arnold D, *Research Design and Statistical Analysis*. *Lawrence Erlbaum*, 2nd ed, p. 508, 2003

[22] Chapelle, O., Ji, S., Liao, C., Velipasaoglu, E., Lai, L., & Wu, S. L. "Intent-based diversification of web search results: metrics and algorithms. " *Information Retrieval*, 2011, vol. 14, no. 6 pp. 572-592.

[23] Sakai, Tetsuya. "Bootstrap-based comparisons of IR metrics for finding one relevant document." In: *Asia Information Retrieval Symposium*. Springer, Berlin, Heidelberg, 2006. p. 374-389.

[24] Mchugh, Mary L. "Interrater reliability: the kappa statistic." *Biochemia medica: Biochemia medica*, 2012, vol. 22, no. 3 pp. 276-282.

[25] Falotico, Rosa, and Piero Quatto. "Fleiss' kappa statistic without paradoxes." *Quality & Quantity*, 2015, vol. 49, no. 2 pp. 463-470.

[26] Ke Wang and Xiaojun Wan. SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 4446-4452.

[27] Stent A., Marge M., Singhai M. (2005) Evaluating Evaluation Methods for Generation in the Presence of Variation. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science, vol 3406. Springer, Berlin, Heidelberg.

[28] Novikova, J., Dusek, O., Curry, A.C., & Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. EMNLP.