An Automatic Evaluation Method for Open-domain Dialogue Based on BLEURT

Shih-Hung Wu
Chaoyang University of Technology
Computer Science and Information Engineering
Taichung, Taiwan
shwu@cyut.edu.tw

Jia-Jun Lee
Chaoyang University of Technology
Computer Science and Information Engineering
Taichung, Taiwan
s11027603@gm.cyut.edu.tw

***Abstract*—Automatic open-domain dialogue generation is an important topic in natural language generation research. Due to the lack of good automatic evaluation methods, it is usually evaluated manually, which makes it difficult t o c ompare d ifferent generation models. Automatic evaluation methods of natural language used in the past often required a reference corpus.**

However, for open-domain dialogue generation, the reference corpus will limit the possibilities of generation. In order to evaluate the quality of a generative model stably, we study the learning to evaluation approach to the generative dialogue based on deep learning method. Experimental corpus includes conversations collected from the web and conversations generated using the GPT-2 model. We manually evaluate these dialogue as a standard, and our system uses the BLEURT-20 model to learn an automatic evaluation model. We find t he m odel i s s uitable a s a n automatic evaluation mechanism for dialogue generation.

***Index Terms*—BLEURT-20, Dialogue generation, evaluation by learning**

I. Introduction

Chinese natural language generation has become stronger and stronger in recent years, and more advanced mechanisms are needed to assess the quality of generated sentences. Previous studies have been based on manual evaluation, which takes a lot of time and manpower to obtain scoring data on sentences. This assessment method is unstable and difficult t o r eproduce. E xisting automatic evaluation models, such as BLEU [1] and ROUGE [2], require reference text. Both are assessment methods that rely on sentence similarity. The difference i s t hat BLEU focuses on calculating accuracy, while ROUGE focuses on calculating recall, and both evaluate scores by calculating n-grams. While it performs well on translation tasks, it has been a weakness in evaluating non-sequence-to-sequence generation types of data, such as questions and answers or conversations.YiSi [3] and BERTscore [4] use embedding markers for sentence context to calculate the similarity between two sentences. Others include BEER [5],RUSE [6]and ESIM [7]uses end-to-end training metrics and relies on a large number of manual evaluation data. Xiaolce [8]uses Conversation-turns Per Session (CPS) to calculate the quality of chats. The method of calculating CPS is to learn a better dialogue pattern from the user's feedback in the conversation, the higher the CPS, the better the learning effect, which in turn improves the quality of the NLG model. The recently launched BLEURT [9], [10] by Google uses the most advanced pre-training methods and fine-tuning with WMT indicator data to make BLEURT models have made major breakthroughs in machine translation.

So we use the BLEURT model as a base training model to evaluate tasks about conversations or questions and answers. The BLEURT model provides a model for training evaluation methods and automatically evaluating dialogue, making the evaluation more accurate in terms of accuracy. For this mission, we added different data to compare, in which we used GPT-2 Chinese [11]. GPT-2 Chinese is a model designed for generating Chinese, and GPT-2 is built on a transformer decoder, the BERT model [12], [13]was built through transformer encoder. The difference is that GPT-2 is like a traditional language model, outputting only one tokenat a time. WoBERT [14]is based on BERT, which mainly generates missing words in sentences, so it is not suitable for this experiment. In this study, we used the experimental method of the BLEURT model to derive a model for evaluating Chinese Traditional conversations. First we'll cover the data we used and the three tasks, followed by the use of two models, and finally the results.

The organization of the paper is as follows: First we' ll cover the data set we used and then describe the experiment design, followed by the results and conclusions.

II. Data Sets

In this study, we used three different s ets o f one-shot open-domain dialogue, as follows.

A. Real World Dialogue

The fist data set is a collection of real world. We select 1,000 dialogue from the PTT-Gossiping-Corpus [15] published on Kaggle.com. Some example of the 1,000 Post-Reply style one-shot dialogue is shown in Table 1.

978-1-6654-6603-5/22/$31.00 ©2022 IEEE
DOI 10.1109/IRI54793.2022.00030

TABLE I
Real world dialogue examples in the PTT-Gossiping-Corpus data set

	Dialogue
post:	右駕和左駕的由來 Origin of right-hand drive and left-hand drive
reply:	美國刻意與英國不同，才出現左駕 The US is deliberately different from the UK, which is why there is a left-hand drive
post:	為什麼吃素不能吃蔥??? Why vegetarians should not eat green onions?
reply:	因為沒寫到，想吃的佛教徒就可以找理由吃了。 Because it's not written, Buddhists who want to eat can find reasons to do so
post:	韓國團體的歌都有 RAP? Korean groups have RAP in their songs?
reply:	韓國每個團的歌都一樣公式啊 Every Korean group has the same song formula
post:	蘭州拉麵好吃嗎 Is Lanzhou Ramen good?
reply:	能夠吃到當場現拉麵條的，台灣有這種店嗎? Is there such a place in Taiwan where you can have freshly made ramen on the spot?
post:	米其林可信度高嗎?? How reliable is Michelin?
reply:	個人吃過覺得還好 Personally, I've had it and I think it's fine.
dialogue:	學生買什麼可以保值？ What students can buy to preserve their value?
reply:	去研究古董，書畫、鐘錶、相機之類的 To study antiques, paintings, clocks and watches, cameras and else.
dialogue:	有沒有一有地震中央氣象局就會被 DDOS Does the Central Weather Bureau get DDOS when there is an earthquake?
reply:	某新聞台表示：這是駭客攻擊 This is a hacking attack, says a news station.
dialogue:	跟風看紅白的人是不是很多? Are there many people who follow the trend to watch the NHK red and white?
reply:	我是來聽演歌的這樣才有節慶的感覺 I'm here to listen to the enka so it feels like a festival.
dialogue:	貓貓早上是不是都不看人？ Do cats not even look at people in the morning?
reply:	這貓貓好可愛 This cat is so cute.
dialogue:	國語女歌手誰排第一？ Who is the number one female Mandarin singer?
reply:	阿妹人氣退？票難買的要命你說人氣退 A-Mei's popularity is fading? It's hard to get tickets and you're saying she's not popular.

B. Generated Dialogue Response

In the second data set, we use a GPT-2 model trained from the PTT-Gossiping-Corpus dataset, which is about 750,000 dialogue pairs, to generate a new set of response to the same post part of the previously selected data set. The response of each post is generated, and obtains a new 1000-dialogue data set; some samples are shown in Table 2. Here we give a brief description on the GPT-2 generation model.

1) GPT-2 model: In recent years, due to the new BERT model, many natural language processing models have also taken turns to perform in the rankings of natural language processing tasks, including BERT models and GPT models [16],Transformer XL models [17],XLNet models [18]–[23]and more. One of them is the GPT-2 model of the OpenAI organization, a new pre-trained model based on the GPT model [24]. We can use the GPT-2 model to write smooth and reasonable articles or answers. The GPT-2 model is based on the output of only one token at a time, and whenever a new word is generated, the word is added after the previously generated word sequence, which becomes the next new input to the model. This mechanism, called auto-regression, is also the main mechanism that makes the RNN model stand out.

The GPT-2 model aims to do supervised tasks using unsupervised pre-trained models. Because of the temporality of text data, an output sequence can be labeled as the product of a series of conditional probabilities (1).

$$p(x) = \prod_{i=1}^{n} p(S_n|S_1, ..., S_{n-1}) \qquad (1)$$

The idea of the equation (1) is the one that is known Text input $input = \{S_1, S_2, ..., S_{n-k-1}\}$ ' , to predict the unknown output $output = \{S_{n-k}, ..., S_k\}$ ' , so the model can be represented is the form of *p(output|input,task)*. In decaNLP [25], the proposed MQAN model can combine machine translation, semantic analysis, relational extraction, natural linguistic reasoning, etc. 10 Class tasks are unified into a single classification task, and there is no need to design a separate model for each subtask. When we use the GPT-2 model, because to avoid the generation of indecent text, word counts that are too short, or a large number of repetitive words in the statement, we adjust the model internally, and if the GPT-2 model generates the above problems, let GPT-2 The model is regenerated once.

C. Human Scoring on the data set

In the third data set, we use 200 of the 1000 dialogue pairs selected from the PTT-Gossiping-Corpus data and label them manually. These data set will be our training and test data sets for the BLEURT model to learn how to score the dialogue.

D. Manual Scoring Methods

In terms of manual scoring, we asked 10 people to rate the generated 1000 dialogue pairs and also the collection of PTT dialogue 1000 pairs in the same standard way. The scoring criteria are divided into 4 levels: 1. The answer is not what was asked. 2. There are dialogue, but too briefly. 3. Answer the question, but the answer is not smooth or polite. 4. The answer matches the question and the sentence is smooth. Each pair are scored separately by manual evaluation from 1 to 4, and a total of 2,000 scoring scores. In terms of calculating scores, we take the problem that if most people in this question are rated 1 and only one or two people have a score of 4, then delete the outliers

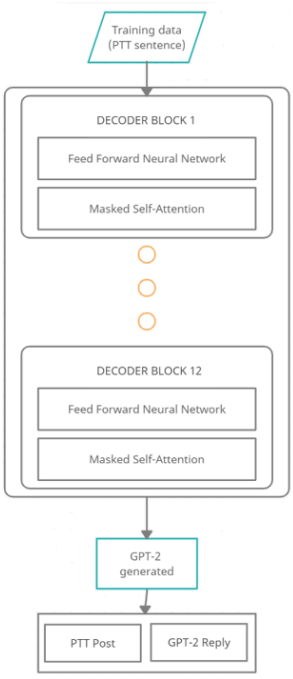

Fig. 1. The flow chart of using GPT-2 model to generate dialogue.

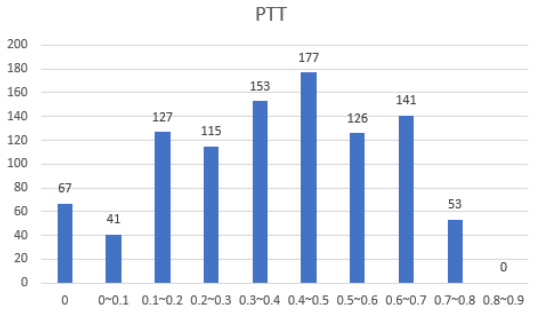

Fig. 2. Score distribution of the manually scoring result on the real world dialogue selected from the PTT-Gossiping-Corpus.

TABLE II
Use GPT-2 to generate respons dialogue with ten dialogue.

	Dialogue
post:	全家便利商店是不是很倒楣 ???
	Is it bad luck for family convenience stores?
reply:	你去全家超商就知道了
	Just go to a family convenience stores and you'll see
post:	律師去簽志願役有沒有搞頭??
	Is it possible for a solicitor to sign up for Voluntary military?
reply:	不然去年輕人家開保時怎麼辛苦?
	Otherwise, how hard would it be to go to a young man's house to drive a Porsche?
post:	手遊抽卡要花多少??
	How much does it cost to draw cards in a handicap tournament?
reply:	而且都沒有要外國人在用才有抽到
	And there is no need for foreigners to use it to get it.
post:	為什麼有人很喜歡幫動物加台詞？
	Why do people like to add lines for animals?
reply:	因為是台灣人
	Because it is Taiwanese
post:	有沒有社交恐懼症的八卦
	Gossip about social phobia
reply:	你應該先擔心吃完薯條就會有新的問題
	You should first worry that you'll have a new problem after eating the fries
post:	希望臺灣人多點同理心
	I hope the Taiwanese will be more empathetic
reply:	台灣的人一定有你們的人權，不然就要完蛋了。
	The people of Taiwan must have your human rights or they will be doomed
post:	鞭炮的功用?
	The function of firecrackers
reply:	鞭炮炸砲炸彈快爆炸了
	Firecrackers, bombs, bombs are about to explode
post:	最強的宵夜貼圖是什麼的八卦
	Gossip about what is the strongest late-night posting
reply:	你是不是又要貼新聞
	Are you going to post the news again?
post:	有沒有早九晚五的工作？
	Do you have a nine-to-five job
reply:	我也是人生沒有這種煩惱？
	I don't have such troubles in my life
post:	躺在床上起不來
	Can't get up from bed
reply:	你真的是不知道什麼
	you really don't know what

and then do the average. Finally, the scores of each pair are compared, we will give 1 if the score of the PTT pair is high, and 0 if the high score is on the GPT-2 pair, see Table 4.

Figure 2 and Figure 3 are the distribution diagrams of each data set after the manual scoring, the number is normalized.

III. Experiments Design

In this study, three experiments are designed to test if BLEURT is suitable for the dialogue evaluation. The first experiment is use the BLEURT model to automatically scores the dialogue generatedby GPT-2 and the dialogue in PTT corpus. The second experiment is comparing the six BLEURT pre-trained models and the model finetuned with the PTT data to observe the consistency of the results of manual evaluation and scoring of the model. The third experiment is to score the dialogue we wrote ourselves, comparing whether the model will give a higher score on the carefully generated corpus.

A. The First Experiment

In the first experiment, based on the post part of 1,000 dialogue pairs selected from the PTT dataset, we use GPT-2 model to generate 1,000 responses. BLEURT model is then used to give score on both the original response and the GPT-2 generated response.

B. The Second Experiment

In the second experiment, we conduct the automatic scoring with six BLEURT models: Bleurt-20, Bleurt-base-512, Bleurt-large-512, Bleurt-tiny-512, Bleurt-fine-tune1

TABLE III
Sample dialogue pairs, where answers is written by our team member, which we considered as better dialogue.

	Dialogue
post:	世界上最難聽的國歌是哪首
	What is the worst national anthem in the world?
reply:	你昨晚唱的那首
	The song you sang last night
post:	正妹再可愛但性格不好也不受歡迎嗎
	Even a cute girl with a bad personality is unwelcome
reply:	性格不好，就算你多可愛都還是不受歡迎吧
	If you have a bad personality, you will not be welcomed even if you are cute
post:	目睹別人分手怎麼辦？
	What to do when you see someone break up
reply:	假裝沒看到
	Pretend you don't see it
post:	有沒有味噌湯是如何做的八卦?
	Have you got any gossip about how miso soup is made?
reply:	網路上很多教學
	Lots of teaching on the web
post:	雨天會影響發文嗎?
	Will the rain affect your posting?
reply:	會吧，我雨天都不喜歡出門的
	Yes, I don't like to go out in the rain
post:	常有測速改成藏有測速照相更可怕？
	Frequent speed test changed to hidden speed camera is even scarier
reply:	看的到國家收入增加的未來
	Seeing the future of increased national income
post:	丹丹開在台北會不會排隊人潮爆滿到整個馬路?
	Will there be a queue for Dandan in Taipei that will fill up the whole road?
reply:	我覺得會，台灣人就是喜歡排隊
	I think so, Taiwanese people like to queue
post:	遇到愛情欸騙子怎麼辦？
	What to do when you encounter a love liar
reply:	只能下次換我騙別人了
	I'll have to lie to someone else next time
post:	關於昨天的海底總動員 3 首映會
	About yesterday's premiere of Finding Nemo 3
reply:	沒看，之後也沒有打算去看
	I didn't see it and I don't plan to see it afterwards
post:	大家有收到地震預警或簡訊嗎
	Have you received an earthquake warning or text message?
reply:	國家邊緣人，沒收過
	I'm National misfit,not received

TABLE IV
Amount of data used and average score after manual scoring.

dataset	quantity	manual scoring
PTT	1000	0.39
GPT-2	1000	0.21
Generate data	200	NO

and Bleurt-fine-tune2. The goal is to find the further finetuning is necessary or not.

C. The Third Experiment

In the third experiment, we used the model to score the high quality dialogue that we prepared manually and see if the models would give higher score than the original PTT dialogue.

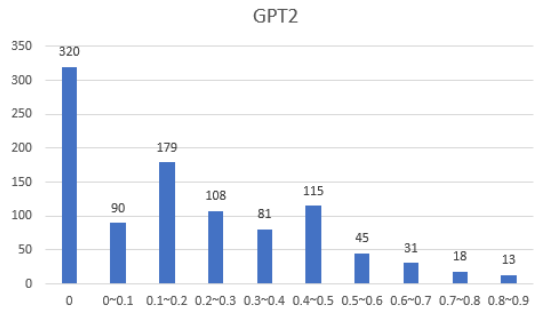

Fig. 3. Score distribution of the manually scoring result on the GPT-2 generate dialogue.

TABLE V
PTT and GPT-2 manually scored dialogue and scores, where 5 data and GPT-2 dialogue are scored, if it is PTT labeled 1 and a GPT-2 statement high is labeled 0, and the results of these six models are observed to match the manual score.

	Dialogue	Score
post:	中文書翻譯成英文書會出名嗎?	
	Will Chinese books become famous when translated into English?	
PTT:	真的想太多了國際學術市場很競爭的	70.8
	You really think too much. The international academic market is very competitive.	
GPT2:	翻成英文拼音比較有梗	33.3
	It's more fun to turn it into English pinyin	
post:	家裡附近的野狗一直叫怎麼辦?	
	What to do if a wild dog keeps barking near your home?	
PTT:	這很難搞，基本上無解	75.0
	This is very difficult and basically insurmountable.	
GPT2:	快去看醫生	50.0
	Go and see the doctor	
post:	誰才是當今跳投最厲害的男人?	
	Who is the best jump shooter today?	
PTT:	歷史第一姆斯	56.2
	The Williams made history.	
GPT2:	當然是林書豪	83.3
	Of course Jeremy Lin	
post:	練啞鈴與伏地挺身哪個 CP 值高?	
	Which has the highest CP value, the dumbbell or the burpee??	
PTT:	不需要啞鈴啦小心練太壯這樣要吃很多	55.0
	You don't need a dumbbell. Be careful of getting too strong, you'll eat a lot.	
GPT2:	你身材好壯.....	0
	You're in great shape.	
post:	希望新的一年大家身體健康	
	Hope everyone is healthy in the new year	
PTT:	祝病魔住你家	37.5
	May the sickness live in your home	
GPT2:	好過日子會好一點	50.0
	It's better to have a better life	

IV. BLEURT Evaluation Model

In this study, we used the BLEURT model as our scoring tool. The BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) model is a model that is based on the results of a research extension of the BERT model, which captures the linguistic characteristics of sentences and is a model that enables automated scoring.

A. The pre-training model

The BLEURT pre-trained model is trained based-on BERT using Chinese Wikipedia. We use the domain corpus to amplify the BLEURT pre-trained model. In our experiments we used the pre-trained model, BLEURT-20, as our Baseline model. In comparison, we use models of different tokens and size, namely base-512, large-512, tiny-512. Finally, on top of BREWUT-20, plus 2000 data obtained after manual scoring, we used the training data to fine-tune the model, and obtained Blend-fine-fine-tune1 and BLEURT-fine-tune2, respectively, with 100,000 and 200,000 training epochs, respectively.

The BLEURT model uses regression or classification loss on pre-trained tasks to lose tasks-level and weights added. τ_k is then defined as the target vector for each task, such as ROUGE's precision, recall rate, and F-score.

If τ_k a regression task,$\boldsymbol{\ell}_2$ loss is used, where τ_k is τ_k on, τ_k, is calculated on the basis of embedding by using a linear layer for a particular :

$$\boldsymbol{\ell}_{pre-training} = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{k} r_k \boldsymbol{\ell}_k(\tau_k^m, \widehat{\tau}_k^m) \qquad (2)$$

In Equation (2), m is the target vector, M is the number of synthetic examples, τ_k The hyperparameter weights obtained through grid search.

The BLEURT model adds BERT to the training model in addition to the 260,000 data manually evaluated by the WMT Metrics task and the news field Transfer learning, using contextual word notation and then using pre-training techniques, effectively improves the accuracy of the BLEURT model.

B. The new pre-training model BLEURT-20

A further improved version of BLEURT model is the BLEURT-20 model [26]. BLEURT is based on the XLM-RoBERTa or mBERT pretrained model, while BLEURT-20 is based on the RemBERT pretrained model. The RemBERT model is a multilingual BERT model and size-down with knowledge distilling. Thus, RemBERT can be used in multilingual applications and keep the size small. Distilling technique has been used successfully to decrease the number of parameter of a deep learning model [27] with the help of more unlabeled data [28], [29]. The BLEURT-20 is distilling with a training data with 13 different languages from Wikipedia (English, German, Chinese, Czech, Russian, Finnish, Estonian, Kazakh, Lithuanian, Gujarati, French, and Turkish). The size of the unlabeled training set is 80M.

C. Transfer the BLEURT-20 Model for Dialogue Evaluation

Our goal is to evaluate the quality of the generated dialogue, which is very different from the original task of BLEURT-20. More fine-tuning is necessary. Table VI shows the results of eleven experiments with different number of epoch. Where the BLEURT-20 model is finetuned with our dialogue evaluation training set : PTT-Gossiping-Corpus. We can find that the best result is from model trained with 20,000 epoch. The experiment shows that the model can be finetuned for the quality dialogue evaluation.

TABLE VI
The consistency the evaluation result between the model with different finetuning epoch and human evaluation.

Pre-train Model	epoch	consistency
finetune-model1	5000	65%
finetune-model2	50000	71%
finetune-model3	100000	86.8%
finetune-model4	150000	74.3%
finetune-model5	200000	90.9%
finetune-model6	250000	90.3%
finetune-model7	300000	82.2%
finetune-model8	350000	43.1%
finetune-model9	400000	73.2%
finetune-model10	450000	71.8%
finetune-model11	500000	72.8%

In this study, we used six different p re-trained models within the BLEURT model, namely Bleurt-20, Bleurt-base-512, Bleurt-large-512, Bleurt-tiny-512, Bleurt-fine-tune1 and Bleurt-fine-tune2 t o s core d ialogue w ith the same post but different response.

In this study, we used six different p re-trained models within the BLEURT model, namely Bleurt-20, Bleurt-base-512, Bleurt-large-512, Bleurt-tiny-512, Bleurt-fine-tune1 and Bleurt-fine-tune2. A nd u sing t hree different answers to the same question, BLEURT had six different pre-trained models evaluate the scores of these three data.

V. Experimental Results

A. Experiment 1 Result

The first experiment result is shown in TABLE VII. The score given by six pre-trained models Bleurt-20, Bleurt-base- 512, Bleurt-large-512, Bleurt-tiny-512, Bleurt-fine-tune1 and Bleurt-fine-tune2. We can find that the fine-tuned model give higher score on the PTT dataset, while the score of GPT-2 is higher than the score of PTT for some model without finetuning. This result shows that finetuning is necessary.

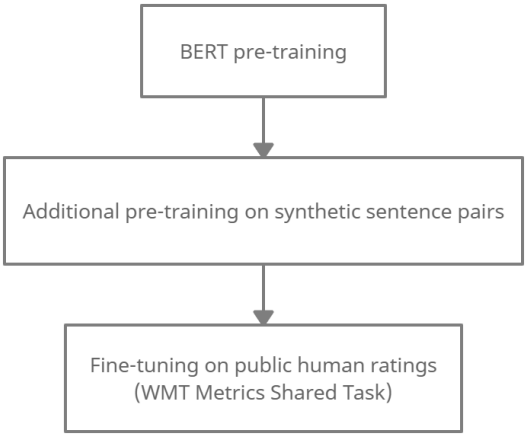

Fig. 4. The flow chart of building the BLEURT model.

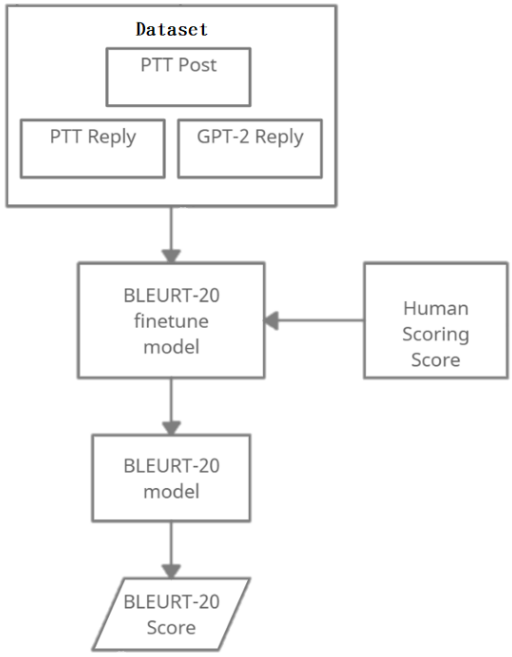

Fig. 5. The flow chart of using the BLEURT model.

TABLE VII
The average score of PTT and GPT-2 on the BLEURT model.

	Bleurt -20	Bleurt -base	Bleurt -large	Bleurt -tiny	Bleurt -fine -tune1	Bleurt -fine tune2
PTT	14.3	16.5	29.1	36.3	36.9	38.1
GPT-2	15.5	15.6	36.4	61.0	29.6	21.3

B. Experiment 2 Result

The second experiment result is shown in TABLE VII. Here we compare the consistency of the BLEURT model with the manual score on which dialogue is a better dialogue. The finetuned model has a higher consistency while the original model give low consistency.

TABLE VIII
Consistency scores for manual scoring and BLEURT six pre-trained model scores.

	Bleurt -20	Bleurt -base	Bleurt -large	Bleurt -tiny	Bleurt -fine -tune1	Bleurt -fine tune2
consistency	54.1	51.3	49.4	38.4	64.9	92.2

C. Experiment 3 Result

The third experiment result is shown in TABLE IX. We ask the BLEURT model to give comparison on our human generated dialogue to the PTT data and to the GPT-2 data. Also, the finetuned model has a higher consistency while the original model give low consistency result.

TABLE IX
In two hundred sentences test set, bleurt model scoring, PTT, and GPT-2 were used to determine the same proportion as manually generated sentence.

	Bleurt -20	Bleurt -base	Bleurt -large	Bleurt -tiny	Bleurt -fine -tune1	Bleurt -fine tune2
human VS PTT	57.5	42.0	48.5	86.0	77.5	85.5
human VS GPT2	57.5	40.5	43.0	58.5	51.5	67.5

VI. Conclusions

In the previous work, we find t hat B LEURT m odel is suitable for the automatic evaluation of generated dialogue to some extent. This paper we find t hat m ore training on the BLEURT-20 model get better consistent raking result than the BLEURT-fine-tune1 p re-trained m odel in previous work. In this study, we used the real world PTT-Gossiping-Corpus dialogue dataset and GPT-2 generated dialogue dataset. We find t hat t he m anually evaluation labelling can help the ranking model to learn how to rank dialogue generation models just like human. The consistency can reach 90

VII. Acknowledgements

This study was supported by the Ministry of Science and Technology under the grant number MOST 110-2221-E-324-011.

References

[1] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J.(2002). Bleu: a method for automatic evaluation of machine translation. 10.

[2] Witten, E. (1989). Quantum Field Theory and the Jones Polynomial. Commun. Math. Phys., 121:351–399. [,233(1988)].

[3] Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H.,and Inkpen, D. (2017). Enhanced lstm for natural language inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver, July.ACL.

[4] Lo, C.-k. (2019). YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available re- sources. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 507–513, Florence, Italy, August. Association for Computational Linguistics.

[5] Shimanaka, H., Kajiwara, T., and Komachi, M.(2018). RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers,pages 751–758, Belgium, Brussels, October. Association for Computational.

[6] Stanojević, M. and Sima' an, K. (2014). BEER:BEtter evaluation as ranking. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 414–419, Baltimore, Maryland, USA, June. Association for Computational Linguist.

[7] Zhang*, T., Kishore*, V., Wu*, F., Weinberger,K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In International Conference on Learning Rep.

[8] Sellam, T., Das, D., and Parikh, A. P. (2020).Bleurt: Learning robust metrics for text generation. In Proceedings of ACL.

[9] Zhou, L., Gao, J., Li, D., and Shum, H.-Y. (2020).The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. Computational Linguistics,46(1):53–93, 03.

[10] Wu, S. and Yeh, C. (2021). Quality evaluation of the general domain chinese dialogue generation models with a bleurt-based model. In 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), pages 431–436, Los Alamitos, CA, USA,aug. IEEE Computer Society.

[11] Devlin, J., Chang, M.-W., Lee, K., and Toutanova,K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.arXiv preprint arXiv:1810.04805.

[12] Devlin, J., Chang, M.-W., Lee, K., and Toutanova,K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.ArXiv, abs/1810.04805.

[13] Du, Z. (2019). Gpt2-chinese: Tools for training gpt2 model in chinese language. https: //github.com/Morizeyao/GPT2-Chinese.

[14] Su, J. (2020). Wobert: Word-based chinese bert model - zhuiyiai. Technical report

[15] Kai-Chou Yang, (2019). Ptt-gossiping-corpus.

[16] Chen, M., Radford, A., Child,R., Wu, J., Jun, H.,Dhariwal, P., Luan, D., and Sutskever, I. (2020).Generative pretraining from pixels.

[17] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q.,and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[18] Burtsev, M. S. and Sapunov, G. V. (2020).Revisiting Pre-Trained Models for Chinese Natural Language Processing, Memory transformer.

[19] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pages 657–668,Online, November. Association for Computational Linguistics.

[20] Parisotto, E., Song, H.F., Rae, J. W., Pascanu, R.,Gulcehre, C., Jayakumar, S. M., Jaderberg, M.,Kaufman, R. L., Clark, A., Noury, S., Botvinick,M. M., Heess, N., and Hadsell, R. (2019). Stabilizing transformers for reinforcement learning.

[21] Shen, Z., Zhang, M., Zhao, H.,Yi, S., and Li, H.(2018). Efficient attention: Attention with linear complexities. CoRR, abs/1812.01243.

[22] Sukhbaatar, S., Grave, E., Lample, G., Jégou,H., and Joulin, A. (2019). Augmenting self-attention with persistent memory. CoRR, abs/1907.01470.

[23] Vecoven, N., Ernst, D., and Drion, G. (2020).A bio-inspired bistable recurrent cell allows for long-lasting memory.

[24] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf,R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q.,and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October. Association for Computational Linguistics.

[25] McCann, B., Keskar, N. S., Xiong, C., and Socher,R. (2018). The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730.

[26] Pu, A., Chung, H. W., Parikh, A. P., Gehrmann,S., and Sellam, T. (2021). Learning compact metrics for mt. In EMN.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In NIPS Deep Learning and Representation Learning Workshop.

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In NeurIPS 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing.

[29] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. arXiv.