Automating Human Evaluation of Dialogue Systems

Sujan Reddy
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India
sujanreddy242@gmail.com

Abstract

Automated metrics to evaluate dialogue systems like BLEU, METEOR, etc., weakly correlate with human judgments. Thus, human evaluation is often used to supplement these metrics for system evaluation. However, human evaluation is time-consuming as well as expensive. This paper provides an alternative approach to human evaluation with respect to three aspects: naturalness, informativeness, and quality in dialogue systems. I propose an approach based on fine-tuning the BERT model with three prediction heads, to predict whether the system-generated output is natural, fluent and informative. I observe that the proposed model achieves an average accuracy of around 77% over these 3 labels. I also design a baseline approach that uses three different BERT models to make the predictions. Based on experimental analysis, I find that using a shared model to compute the three labels performs better than three separate models.

1 Introduction

The evaluation of Natural Language Generation (NLG) systems has generally been carried out by using automatic metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), etc. However, previous work Novikova et al. (2017) demonstrated that these metrics only weakly reflect human judgments of these NLG systems' output, and some form of human evaluation is required to better measure the quality of such NLG systems. Human annotators are generally asked three questions to evaluate whether a system-generated reference is acceptable or not . These questions, along with the corresponding aspects, are:

1. **Naturalness**- Could the utterance have been produced by a native speaker?
2. **Quality**- Is the utterance grammatically correct and fluent?
3. **Informativeness**- Does the utterance provide all the useful information from the meaning representation?

Since human evaluations can be expensive and time-consuming, an automated approach to flag such instances could make it easier for system designers to garner insights into the kind of instances the system is failing to generate good text for. In this paper, I propose a BERT-based model trained to predict answers to questions pertaining to the three aspects: naturalness, quality, and informativeness, with a "YES" (label=1) or a "NO" (label=0). The proposed model automatically flags system-generated references that are not up to a predefined standard. To the best of my knowledge, this is the first attempt to develop an automated model for predicting scores pertaining to multiple aspects of a system-generated reference.

The major contributions of this work can be summarized as follows : First, I propose a binarization scheme to binarize the human judgment scores in the dataset as these scores tend to be very subjective. A threshold is set, and all scores above the threshold are assigned a label and the scores below the threshold are assigned another label. Second, the BERT-based model is fine-tuned to predict three labels, answering the questions corresponding to the three aspects of the system-generated reference. I also perform an ablation study where three separate BERT-models are trained independently, each of which predicts a label.

The remainder of this paper is structured as follows: Section 2 talks about the recent works in the same domain. Section 3 discusses about the BERT-model that is used for the experiments. In section 4, I discuss about the dataset, pre-processing required, hyper-parameters as well as the baseline model's design. In section 5, I discuss about the performance of the proposed apporach in comparison with the baseline model. Finally in 6, I draw conclusions and outline future works.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 229 - 234
July 10-15, 2022

2 Related Work

Several works have been proposed in recent years which focus on fine-tuning BERT (Devlin et al., 2018) and its variants to evaluate the quality of a system-generated text. These approaches tend to correlate much better with human assessments. BERTScore (Zhang et al., 2019) compares the similarity of each token in the system generated reference with each token in the original reference using contextual embeddings rather than exact matching. This metric was observed to relate very closely to human judgments for image captioning systems. MoverScore (Zhao et al., 2019) is another metric that combines contextualized representations with distance measures. This metric was observed to generalize well across various tasks like summarization, machine translation, image captioning, and data-to-text generation. BLEURT (Sellam et al., 2020) is a BERT-based model that was pretrained on a large amount of synthetic data. This model can then be fine-tuned on a relatively small number of human judgments. It was observed to be very effective when the training data is scarce and imbalanced. COMET (Rei et al., 2020) is another neural framework that is used for training multi-lingual machine translation quality evaluation models.

All these works evaluate only the quality of the system-generated reference. While quality is correlated with the other aspects of the utterance, it might not be sufficient to capture all insights about an incorrectly generated text with just a single aspect. A grammatically correct text could lack some vital information that was present in the original reference (informativeness) or may not capture the natural speech patterns of a native speaker (naturalness).

Liu et al. (2021) proposed an automatic method for evaluating the naturalness of generated text in dialogue systems by fine-tuning a BERT-based model. The proposed model predicts a score between 1 and 6, indicating how natural the system-generated utterance is. However, this work does not consider that human judgments tend to be subjective. The data being fed to the model is therefore ambiguous in nature. In addition, the best model proposed in this paper uses human judgments on other related aspects like quality and informativeness by leveraging the positive correlation between these three aspects. However, this paper proposes a solution that eliminates human evaluation at inference time. Human annotations are used only for training the model. After training, the model can mimic/replace human annotators. Given the success of BERT-based models for system evaluation, I also use pre-trained BERT in my approach.

3 Method

BERT stands for Bidirectional Encoder Representation Transformer. The architecture of BERT was based on the encoder part of Transformers (Vaswani et al., 2017). BERT uses attention mechanism (Bahdanau et al., 2014) to convert the input representation into a better representation that takes context into account (Devlin et al., 2018). BERT makes use of fine-tuning to leverage the knowledge gained from pre-training. This means that BERT is pretrained on a relatively generic task, and the same architecture is fine-tuned on similar downstream tasks.

In this paper, I use the uncased BERT-Base model. that consists of 12 layers, 768 hidden states, and 12 attention heads. We will be leveraging the pre-training knowledge gained from NSP more than MLM. A [CLS] token is added to the system-generated reference's beginning. A [SEP] token is then added to the system-generated reference, followed by the original human-written reference. This is again followed by a [SEP] token. The tokens fed as input are tokenized using WordPiece embeddings. Sequence embeddings are also passed as input which stores information about which sentence the token belongs to. Positional embeddings from the Transfomer model are added to the input word embeddings along with sequence embeddings. So the model takes in two sentences as the input and predicts whether the second sentence follows the first sentence or not. The encoded representation of the [CLS] token contains information about the representation of the entire sequence. This is called pooled output. The pooled output is passed through a linear layer which is then followed by the output layer with 3 nodes having sigmoid activation. The final output is three values indicating the probability that the system reference is natural, fluent, and informative, respectively (see Fig 1).

4 Experimental Setup

4.1 Dataset

I consider the "Human Ratings of Natural Language Generation Outputs" (Novikova et al., 2017)

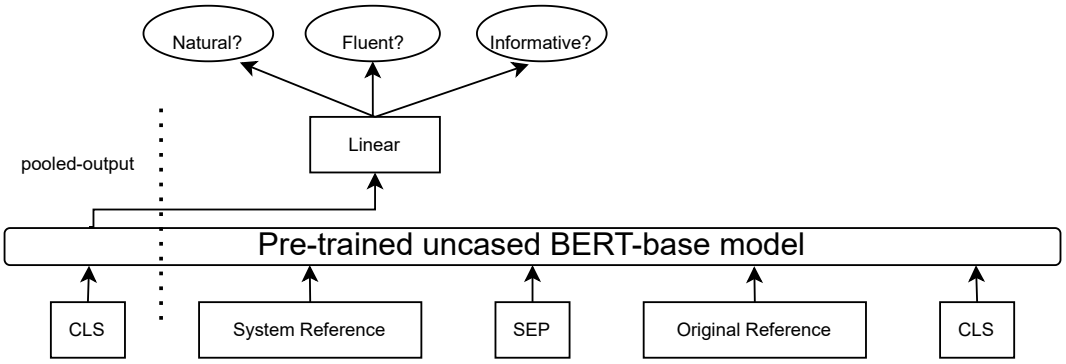

Figure 1: Fine-tuning BERT architecture

dataset in this paper. The dataset contains textual dialogue response from RNNLG[1], TGen[2] and LOLs[3]. These are data-driven natural language generation systems that were applied on 3 different but closely related domains- **SF hotel, SF restaurant** (Wen et al., 2015) and **BAGEL** (Mairesse et al., 2010) respectively. SF hotel and SF restaurant are based on information regarding hotels and restaurants in San Francisco, while BAGEL has information about restaurants in Cambridge. For every NLG system-generated reference, there is also a human-written reference in the dataset. The dataset also contains scores from 3 different human annotators for the system-generated reference's naturalness, quality, and informativeness. These scores were provided on the 6-point Likert-Scale with the lowest score being one and the highest being six. Table 1 contains an example of an instance from the dataset. Here judge refers to the label of the human annotators. Since there are three human annotators, the three labels are 1,2, and 3. The table also presents the BLEU , rouge-L and Meteor scores for the system generated output. These metrics are on higher side, which might indicate that the system-generated output is good. However, the human judge allots low scores for all three aspects for this instance.

Table 2 contains the distribution of the median of the scores from the three annotators over 11,122 instances. For some instances, I observed that more than one NLG system generated the same text. In such cases, the median of all such scores obtained from different NLG systems over the three human judges was considered. If the median is not a whole number, I consider the ceiling of the median score. The higher scores are due to the fact that the dataset considers state-of-the-art NLG systems.

Scores	**naturalness**	**quality**	**informativeness**
1	426	403	153
2	348	501	405
3	801	1071	320
4	1876	1930	1040
5	3383	3531	3427
6	4288	3686	5777

Table 2: Distribution of the median scores

Human annotations on naturalness, quality, and informativeness tend to be subjective. In fact, all three human annotators give the same naturalness score for only 1351 instances, identical quality scores for 1180 instances, and identical informativeness scores for 1772 instances.

Hence, to remove this ambiguity in the dataset, I decided to binarize the dataset by defining a fixed threshold. Novikova et al. (2017) classify all the ratings with scores greater than or equal to 5 as good ratings. Hence, I chose 5 as the threshold. All the instances with median scores below five are assigned a label of '0' and are considered bad utterances. All instances with median scores greater than or equal to 5 are assigned a label of '1' and considered good utterances.

Class	**naturalness**	**quality**	**informativeness**
0	3450	3904	1920
1	7672	7218	9202

Table 3: Distribution of the binarized scores

[1]https://github.com/shawnwun/RNNLG
[2]https://github.com/UFAL-DSG/tgen
[3]https://github.com/glampouras/JLOLS

Field	Value
System Generated Output	x is a french and restaurant near x..
Original Reference	x is a restaurant serving french food, near x
Judge	3
Informativeness	2
Naturalness	2
Quality	2
BLEU-1	0.875
BLEU-2	0.790569415
BLEU-3	0.678604404
BLEU-4	0.5
rouge-L	0.944690265
meteor	0.540778542

Table 1: Example of an instance from the dataset.

Table 3 shows the new distribution after binarizing the dataset. It can be observed that judgments are still skewed towards the higher scores. The ratios of positive (greater than or equal to 5-points) for the three aspects are 68:32, 64:36, and 82:18, respectively. The dataset (11122 instances) is randomly split into train, validation, and test with an 80:10:10 ratio.

4.2 Baseline

To test the performance of the proposed architecture, I use another approach that involves fine-tuning BERT. In this approach, I use three different BERT models, each fine-tuned to predict one of the aspects pertaining to the system-generated utterance. This approach is computationally less efficient than my proposed approach because it takes more time to train and get inferences from 3 different models. Also, this approach utilizes close to three times the memory used by my approach.

4.3 Experimental Setting

The BERT-base model contains 12 layers, 768 hidden states, and 12 attention heads. The pooled output is fed to a linear layer that contains 768 nodes. For all the experiments, I set the batch size to 16. I use the Adam optimizer and set the learning rate to 3e-4. All the models were run for five epochs. Since I use the BERT-Base model, the linear layer has dimension 768.

To deal with the class imbalance problem, I use the balanced cross-entropy function (L) (see equation 1) where $\hat{y}$ refers to the model output and y refers to the ground truth.

$$L = -\beta y log(\hat{y}) - (1-\beta)(1-y)(log(1-\hat{y})) \quad (1)$$

This loss function penalizes the model by a greater factor when it misclassifies an instance with a negative label than a positive label. I tune the parameter β using grid-search for the approach that uses 3 different BERT models. Zhou et al. (2017) suggests utilizing the ratio of negative instances to the total number instances as this factor. So I perform a grid search over values 5%, 10%, 15%, 20%, and 25% lesser as well as greater than this ratio. I observe that the optimal parameters obtained from grid search to be 0.3535, 0.3130, and 0.1454 for naturalness, quality, and informativeness, respectively. I use the same parameter for my approach with a single BERT model with three prediction heads.

5 Results and Discussion

Table 4 reports the comparison of the accuracies between both of my approaches. Given that the data is imbalanced, I also compare the f-1 scores of both of my approaches in Table 5. The tables report the mean and standard deviation of each metric computed over five iterations, each iteration having a different random seed. In Table 4 and Table 5, 3-BERT indicates three separate BERT models, and shared-BERT indicates a single BERT model with three prediction heads. The accuracy and f-1 scores suggest that shared-BERT outperforms 3-BERTs with respect to both measures for naturalness and informativeness. Since the prediction for all three aspects would require similarly encoded input representations, having a shared model instead of 3 individual models can significantly reduce the memory needed. Shared weights act as a regularizer and lessen the chances of over-fitting.

Aspect	3-BERT	shared-BERT
naturalness	76.19 (±1.00)	**77.98*** (±1.99)
quality	**67.66** (±3.14)	66.01 (±1.69)
informativeness	86.48 (±2.31)	**89.04*** (±0.79)

Table 4: Comparison of accuracies of predicting labels for system evaluation. * indicates that the difference is statistically significant with $p < 0.05$.

Aspect	3-BERT	shared-BERT
naturalness	81.81 (±1.60)	**84.63*** (±1.44)
quality	**73.87** (±3.27)	73.17 (±1.90)
informativeness	91.78 (±1.55)	**93.53*** (±0.48)

Table 5: Comparison of f-1 scores of predicting labels for system evaluation. * indicates that the difference is statistically significant with $p < 0.05$

Also, such a model can generalize well on new aspects that can be added in the future. The results suggest that both models can model the data well despite the class imbalance. This can be attributed to the balanced cross-entropy loss function.

I use the ANOVA test (Girden, 1992) to test the statistical significance of the difference in the f-1 scores and accuracies between both approaches. I set the significance level to 0.05. I observe that the results are statistically significant for naturalness and informativeness, which clearly demonstrates that the shared BERT model outperforms the 3-BERT model on these two aspects. For the aspect of quality, 3-BERT shows better performance. However, the gain in performance is not statistically significant. Further, in terms of model complexity, shared-BERT has only 2304 (768x3) learnable parameters more than a single BERT model, and the 3-BERT approach has three times the number of learnable parameters compared to a single BERT model. Hence, shared-BERT is a more efficient model in terms of memory occupied and computational complexity.

Qualitative example: I consider the example instance from Table 1. The scores from the automated evaluation metrics suggest that the system-generated output is a good one. However, the human annotator assigned low scores for this instance. Table 6 presents the scores obtained from both my approaches for this instance. These low probabilities indicate that the system-generated output is not natural, not informative and not fluent. This is an example of an instance which demonstrates the significance of having human annotations, and how the proposed models can mimic human annotators.

Aspect	3-BERT	shared-BERT
naturalness	0.15	0.12
quality	0.28	0.19
informativeness	0.33	0.22

Table 6: Model Output for considered example

6 Conclusion and Future Work

In this paper, I proposed an automated approach to evaluate three aspects of a system-generated sentence : naturalness, quality, and informativeness. I experiment with two BERT-based model approaches. Experimental validation suggests that the proposed approach that uses a single BERT model with three prediction heads is more efficient than three different BERT models with a single prediction head each.

The goal of this paper is to reduce the load on human annotators and automate the evaluation of dialogue systems. I hope that this work will motivate researchers to realize that this process can be automated and be made more reliable with the collection of additional relevant data. Further, aspects other than the three considered in this paper can yield some more insights into the performance of a dialogue system. As an extension of this work, I will verify the performance of my approach on other NLG systems like image captioning, question answering, machine translation, etc.

References

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ellen R Girden. 1992. *ANOVA: Repeated measures*. 84. sage.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ye Liu, Wolfgang Maier, Wolfgang Minker, and Stefan Ultes. 2021. Naturalness evaluation of natural language generation in task-oriented dialogues using bert. *arXiv preprint arXiv:2109.02938*.

François Mairesse, Milica Gasic, Filip Jurcicek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.