# Language Model Transformers as Evaluators for Open-domain Dialogues

**Rostislav Nedelchev**[1] **Jens Lehmann**[1,2] **Ricardo Usbeck**[1,2]
[1]Smart Data Analytics Group, University of Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin and Dresden, Germany
rostislav.nedelchev@uni-bonn.de, jens.lehmann@cs.uni-bonn.de
ricardo.usbeck@iais.fraunhofer.de

## Abstract

Computer-based systems for communication with humans are a cornerstone of AI research since the 1950s. So far, the most effective way to assess the quality of the dialogues produced by these systems is to use resource-intensive manual labor instead of automated means. In this work, we investigate whether language models (LM) based on transformer neural networks can indicate the quality of a conversation. In a general sense, language models are methods that learn to predict one or more words based on an already given context. Due to their unsupervised nature, they are candidates for efficient, automatic indication of dialogue quality. We demonstrate that human evaluators have a positive correlation between the output of the language models and scores. We also provide some insights into their behavior and inner-working in a conversational context.

## 1 Introduction

Lately, deep learning conversational systems have seen increasing interest from industry and academia alike (Chen et al., 2017). These systems find usage in various contexts, starting from personal speech assistants like Google Assistant through the "chatbots" on instant messaging platforms like Facebook Messenger, and finally, conversational services like LUIS[1]. Many of these applications serve the objective of completing a specific function like purchasing a product or booking services (e.g., hotels, flights). Nonetheless, these applications can still profit from open-domain dialogue skills like chit-chatting, which would provide a more human-like interaction with users.

Presently, scientists and engineers working on computer-based conversational systems need human-based evaluation to assess the quality and usability of their work (Dinan et al., 2019; Logacheva et al., 2018; Yoshino et al., 2019). These evaluations are costly in terms of resources. Thus, the field of dialogue systems could take advantage of an automated method for assessing conversations.

Seminal works in text summarization and machine translation have already proposed their field-specific metrics for automated assessments - for the former ROUGE (Lin, 2004), and, for the latter, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). Dialogue system research (Ritter et al., 2011; Serban et al., 2016; Yoshino et al., 2019) constantly uses these metrics. However, Liu et al. show that these metrics based on word-overlap between prediction and references are not reliable for evaluating the usefulness of dialogue systems (2016). Hence, the field should use more sophisticated methods that consider the previous utterances of a conversation and its semantic meaning.

When human annotators evaluate a dialogue, they do not use an explicit reference or necessarily seek word overlap between context and response (or the lack of it). Their assessment bases itself on experience with the language and the implicit knowledge they have about it. The core principle of statistical language models (LM) is to capture and reproduce these properties. LM have proven themselves invaluable in state-of-the-art approaches in natural language processing, and natural language understanding (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019).



[1]https://www.luis.ai/



*Proceedings of the 28th International Conference on Computational Linguistics*, pages 6797–6808
Barcelona, Spain (Online), December 8-13, 2020

Thus, the main aim of this work[2] is to investigate their usability as means for evaluating dialogues since they do not need a reference or supervision. We demonstrate that there is a significant positive correlation between the predictions of language models and human evaluation scores. Furthermore, we provide insights into the inner-workings and behavior of language models in the dialogue context.

## 2 Related Work

In this section, we present earlier work that focuses on dialogue evaluation. Furthermore, we provide a concise introduction to language model transformers and recent advances in this particular set of approaches.

### 2.1 Dialogue Evaluation

Lowe et al. present a cornerstone work in dialogue evaluation. (2017). They propose an *automatic dialogue evaluation model (ADEM)* that employs a neural network approach that approximates human judgment using scored dialogues together with the context, reference response, and one generated by a dialogue system. Reference responses and human annotation scores are hard to obtain. That is, it is challenging to employ the approach on large dialogue datasets. Another cornerstone is the work of Tao et al. (2018), a *Referenced metric, and Unreferenced metric Blended Evaluation Routine (RUBER)*. They suggest a method consisting of two elements: The first one captures the resemblance between a generated and reference response using word vector pooling. The second one uses a neural network to estimate the relevance of a reply. The model is trained to distinguish whether an answer in a dialogue is the original one or a random one from another conversation. A drawback of both approaches above is that they use reference responses to derive a score. Furthermore, Sai et al. (Sai et al., 2019) demonstrate that machine learning approaches for dialogue evaluation like ADEM are susceptible to adversarial attacks.

Other works focus on addressing the issue that there is more than one possible response for a given dialogue context by considering multiple reference responses. For example, Galley et al. (2015) suggest an augmented version of BLEU that uses synthetically generated responses. The algorithm in Sordoni et al. (2015) operates similarly. Sugiyama et al (2019) develop a Support Vector Regression approach to consider multiple references. Gupta et al. (2019) investigate a framework of dialogue-modeling methods combined with a variety of metrics, where they evaluate dialogues using various references.

Zhang et al. (2020) propose BERTscore to calculate text similarity using contextual embeddings. Their work can be used for evaluating text generation against a reference. Unfortunately, it offers no way to evaluate dialogues without a specified ground truth. On another note, Kann et al. (2018) suggest a sentence level fluency metric derived from the perplexity score of a language model given a sentence without involving any references. Their results demonstrate significant positive correlations with human annotators. Nedelchev et al. (2020) experiment with an anomaly detection approach where erroneous dialogues are seen as anomalies.

### 2.2 Language Models

The first application of n-gram-based language models is recorded in the mid-1970s by two independent works of Jelinek (1976) and Baker (1975). Given a sequence of tokens, $T = \{t_1, ..., t_N\}$, a forward language model computes the probability of the sequence by modeling the probability of a token $t_K$ ($K \leq$ N) which has a history up to the $K$-th token (Peters et al., 2018). Some of the initial neural network models (Melis et al., 2017) use initially a context-independent vector representation for a token, which all pass through one or more LSTM layers (Hochreiter and Schmidhuber, 1997). In the end, they produce a context-dependent vector that serves as input to a softmax layer to predict the next token. In a reversed fashion, backward LM use the context to the right of the target token to predict it. In contrast, bi-directional language models use a combination of both to predict the target word.

Radford et al. (2019) propose generative pre-training (GPT2), where they use the transformer (Vaswani et al., 2017) as a forward language model, due to its superiority in terms of long-term memory when

[2]Code and resources to reproduce the results are available on the following link:
`https://github.com/SmartDataAnalytics/transformers_dialogue_evaluators`

contrasted to recurrent neural networks like LSTMs.

Furthermore, Devlin et al. (2019) suggest an innovative way to train language models, also utilizing transformers, specifically Bidirectional Encoder Representations from Transformers (BERT). They invent the masked language model (MLM) where a random subset of tokens from a sequence is masked or replaced, which the model then predicts by using the remaining original context. Furthermore, BERT uses an additional LM objective: next sentence prediction (NSP). It works by teaching a model to recognize whether two sentences appear sequentially in a corpus or not.

Yet another innovative transformer-based language model is XLNet by Yang et al (2019). It combines the best features of a generative LM like GPT2 and a masked LM like BERT by proposing to use the permutations of all factorization orders of a sequence to train. Thanks to it, XLNet learns to utilize knowledge from both sides of the target token, but also the respective context of other positions. Golovanov et al. (2019) demonstrate that pre-trained transformer language models provide benefits for conversational agents.

For completeness, we mention other language models below that utilize transformers but are not integral to this work. We do not employ them in this work because the architectures discussed above already supersede them, or we deem their additions as not adequate for modeling dialogues.

Dai et al. (2019) propose Transformer-XL, a new approach that allows transformers to model even longer sequences by caching and reusing intermediate hidden states. XLNet also utilizes the method in its implementation. Cross-lingual Language Model, by Lample and Conneau (2019), introduces Translation Language Modeling, i.e., randomly masks words in parallel sequences in two languages to teach the model leveraging multi-lingual context. Liu et al. (2019) present Robustly optimized BERT, by just dropping BERTs next sentence prediction and a few other modifications in training. Raffel et al. (2019) introduce the Text-to-Text Transfer Transformer, where the language-modeling objective is using a text-to-text perspective. Conditional Transformer Language model, by Keskar et al. (2019), incorporates conditioning on control codes to guide the generation of tokens.

Besides capturing syntax, LM are also capable to model semantics of sentences. The results of Tenney et al. (Tenney et al., 2019) suggest that contextual word embeddings can encode both syntax and semantics on a sub-sentence level. Furthermore, Zhou et al. (Zhou et al., 2019) conduct a systematic benchmark to evaluate seven LM for their commonsense knowledge and reasoning. Their work suggests that they have a certain degree of those abilities. Commonsense is what would also help for evaluating open-domain dialogues.

## 3 Methodology

In this section, we report on the used datasets for assessing the usability of transformer language models for evaluating dialogue quality, introduce the used approaches in greater detail and describe their relevance to the task at hand.

### 3.1 Datasets

We use the data gathered during the ConvAI1[3] (Burtsev et al., 2018; Logacheva et al., 2018) and ConvAI2[4] (Zhang et al., 2018; Dinan et al., 2019) challenges. The organizers invited competitors to develop dialogue systems that had to address specific tasks. For ConvAI1, the participating systems needed to be able to converse about a topic. In the other competition, the chatbots had to engage in a small-talk while impersonating a pre-defined personality profile (”persona”). In both cases, human annotators evaluated the capability of the dialogue systems to converse by interacting with them and giving a score at the end. For both competitions, the scoring is on dialogue level. In Table 1 and in Figure 1, we present some additional details about the data. However, we do not evaluate the two challenges specifically (topic discussion and role acting). Instead, we aim at general open-domain dialogue evaluation, which implies relevance, coherence, and fluency of the utterances.

[3] http://convai.io/2017/data/

[4] http://convai.io/data/

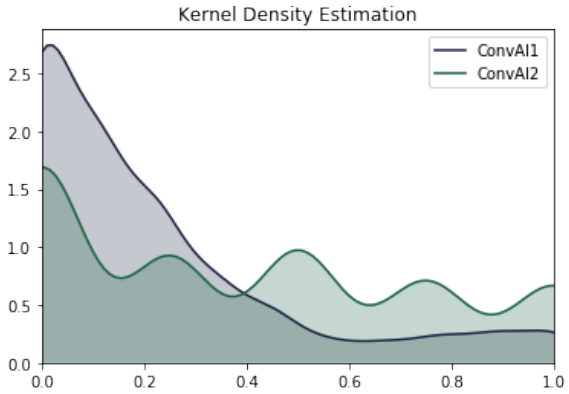



Figure 1: Kernal density estimation of the distribution of annotator scores of the dialogues in ConvAI1 and ConvAI2. We see that the majority of dialogues are evaluated as low quality.

| Feature | ConvAI1 | ConvAI2 |
|---|---|---|
| # Dialogues | 2154 | 2237 |
| Avg # Utterances | 13.9 | 18.1 |
| Avg # Words per Utterance | 7.3 | 8.2 |
| Task | Topic discussion | Role acting |

Table 1: Key features of the dialogue datasets. Only dialogues with three or more utterances were considered as part of this work. From our point of view, a dialogue with two turns cannot reflect the semantic complexities of language.

### 3.2 Language Model Evaluators

In Section 2.2, we presented a concise introduction into transformer-based language models. In the current subsection, we will provide more details about three of those architectures, and how we use them for conducting this study. Our main goal is to use the LM to assign a probability to the utterances in a conversation. We used HuggingFace's Transformers[5] (Wolf et al., 2019) for implementation and pre-trained weights of transformer-based language models.

Since intuition dictates that responses are dependent on their preceding context, we condition the target reply on its history to measure its relevance. Kann et al. (2018) showed how language models could serve as good sentence-level fluency indicators. Thus, the calculated probability from the transformer-based LM can serve as a combined score for fluency and coherency. The following LM are used in this work:

1.) As previously mentioned, BERT (Devlin et al., 2019) is using two language modeling objectives: masked language modeling (MLM) and next sentence prediction. MLM provides no viable way for computing the probability of a target response because it originally substitutes only a random subset of tokens. Thus, there is no consistent and deterministic way to use masked language modeling for assigning a probability score to a response given its context. However, BERT's next sentence prediction is an excellent approach for the current task. It can judge if an utterance is the next one given its contextual predecessor. Thus, we pair up the sequentially appearing sequences in a conversation and compute a probability score for the second reply:

$$P(u_2|u_1) = P(t_{21}, t_{22}, ..., t_{2n}|t_{11}, t_{12}, ..., t_{1m}) \tag{1}$$

,

where $P(u_2|u_1)$ is the probability score of the target response, while $(t_{11}, t_{12}, ..., t_{1m})$ and $(t_{21}, t_{22}, ..., t_{2n})$ are the tokens belonging to the query and response utterances prospectively.

2.) The approach of GPT2 (Radford et al., 2019) is the standard language model approach that factorizes the joint probability over the sequence tokens $(t_1, t_2, ..., t_n)$ as a product of the conditional probabilities (Peters et al., 2018):

$$P(x) = \prod_{i=1}^{i} P(t_n|t_1, t_2, ..., t_{i-1}) \tag{2}$$

In our problem domain, we need to consider two consecutive sequences and capture the coherence between them. Thus, we concatenate them into one, where the context appears first and is then followed by the second utterance. We then compute the joint probability for the second part conditioned on the past:

[5] https://github.com/huggingface/transformers

$$P(x) = \prod_{i=m+1}^{m+n} P(t_{m+n}|t_i, t_{n+1}, ..., t_{m+n-1}) \quad (3)$$

where $m$ is the length of the context, and $n$ is the length of the target utterance.

3.) XLNet (Yang et al., 2019) follows the same general language model approach as GPT2, however, with some additions to its training objective and neural network architecture. First of all, unlike GPT2, XLNet optimizes the model over a sequence w.r.t. all possible permutations of the factorization orders rather than each one separately. Secondly, compared to conventional neural transformers, XLNet adds one more attention stream that includes the positional information of the target token but excluding the content to maintain the autoregressive properties. To compute probabilities for the utterances, we follow the same procedure as described above for GPT2.

In this work, we use a set of hyper-parameter configurations for each of the three language models. We present them in Table 2.

| Name | Details |
|---|---|
| **bert-base-uncased** | 12-layer, 768-hidden, 12-heads BooksCorpus English Wikipedia |
| **bert-large-uncased** | 24-layer, 1024-hidden, 16-heads BooksCorpus & English Wikipedia |
| **gpt2** | 12-layer, 768-hidden, 12-heads news, Wikipedia, fiction books |
| **gpt2-medium** | 24-layer, 1024-hidden, 16-heads news, Wikipedia, fiction books |
| **gpt2-large** | 36-layer, 1280-hidden, 20-heads news, Wikipedia, fiction books |
| **xlnet-base-cased** | 12-layer, 768-hidden, 12-heads same as BERT + news |
| **xlnet-large-cased** | 24-layer, 1024-hidden, 16-heads same as BERT + news |

Table 2: Hyper-parameter configurations (number of layers, size of the hidden state, number of attention heads) of the models and used corpora to pre-train them. Source: `https://huggingface.co/transformers/pretrained_models.html`

### 3.3 Scoring

In Equations 2 and 3, we showed how language models compute a probability score for a whole sequence. However, as an aggregated score over the tokens, it is losing the initial probabilistic distribution over the tokens. Furthermore, since we are dealing with dialogues, i.e., a sequence of utterances, we need to perform two levels of aggregation. The first level is an aggregation of the word tokens within an utterance, while the second is the done while aggregating over the utterances.

Thus, we investigate other possible ways to derive an aggregated score over the word tokens and over the utterances within a dialogue. Besides a product of probabilities, we also look into a sum and an unweighted average, which capture the length of the sequences (utterance or dialogue), which might prove beneficial for a correlation study with human annotators. We normalize all of the scores such that they range between 0.0 (population minimum) and 1.0 (population maximum).

For GPT2 and XLNet, our experiments show that the following formulation correlates the highest with human annotator scores:

$$lm\_dialog\_score = \sum_{u=1}^{Utterances} \left( \frac{\sum_{w=1}^{Words} P_{(w=w)}}{\#Words} \right) \quad (4)$$

We investigated other means to compute an aggregated score on the dialogue level. We present the other results with low correlation coefficients and significance values in the appendix.

### 3.4 Baseline

We take RUBER from Tao et al (2018) as a baseline. The approach initially employs two components that perform two functions. The first one is to calculate a resemblance score using word vector pooling

and references. We aim for an unreferenced evaluation approach akin to a human evaluator. Thus, we use only the second component of the method. This second component can calculate a relevance score for a given response based on its preceding context. It uses a bidirectional GRU network and negative sampling. To reproduce as best as possible the original results of RUBER, we sample 1,449,218 pairs of sequential utterances from the OpenSubtitles dataset (Lison and Tiedemann, 2016).

## 4 Evaluation

In this part of the paper, we will conduct a correlation analysis between the calculated probabilities from the LM and the scores given to dialogues by human evaluators. We provide a closer look at some auxiliary model outputs as well.

### 4.1 Quantitative Assessment

In Table 3, we report the noteworthy Pearson's and Spearman's correlation coefficients between the aggregated probability scores and the evaluations of the dialogues.

The immediate observation of using language models as dialogue evaluators shows that there are gaps in terms of performance between the three different approaches. Most evident is the difference between BERT and the others. Its next sentence prediction objective explains this behavior. Unlike the other two, BERT takes the most structured approach to modeling two sequences. It recognizes the two utterances as separate and captures their information as a whole. Thus, when we compare it to GPT2 and XLNet, it has the advantage of not needing score aggregation on utterance level, because it produces a probability for the whole sentence rather than word for word.

Also, there is a smaller difference in performance between GPT and XLNet. First of all, they share a core foundation as autoregressive language models, thus are more similar to each other than BERT, which also explains their overall behavioral similarity. However, XLNet has a structural improvement in its architecture. Unlike GPT2, it also encodes the positional information of the target token. Thus, similarly to BERT, it can capture more information about a sequence and consequently have a better correlation score.

| **Dataset** | **ConvAI1** | | **ConvAI2** | |
|---|---|---|---|---|
| | $r$ | $\rho$ | $r$ | $\rho$ |
| **bert-nsp-d-sum** | 0.169 | 0.273 | 0.205 | 0.490 |
| **bert-large-nsp-d-sum** | 0.172 | 0.277 | 0.205 | 0.485 |
| **gpt2-u-avg-d-sum** | -0.027 | 0.068 | 0.152 | 0.323 |
| **gpt2-md-u-avg-d-sum** | -0.005 | 0.069 | 0.144 | 0.325 |
| **gpt2-lg-u-avg-d-sum** | -0.038 | 0.048 | 0.127 | 0.325 |
| **xlnet-u-avg-d-sum** | 0.068 | 0.157 | 0.206 | 0.435 |
| **xlnet-lg-u-avg-d-sum** | 0.087 | 0.169 | 0.225 | 0.437 |
| **RUBER-U** | 0.154 | 0.129 | 0.013 | -0.005 |

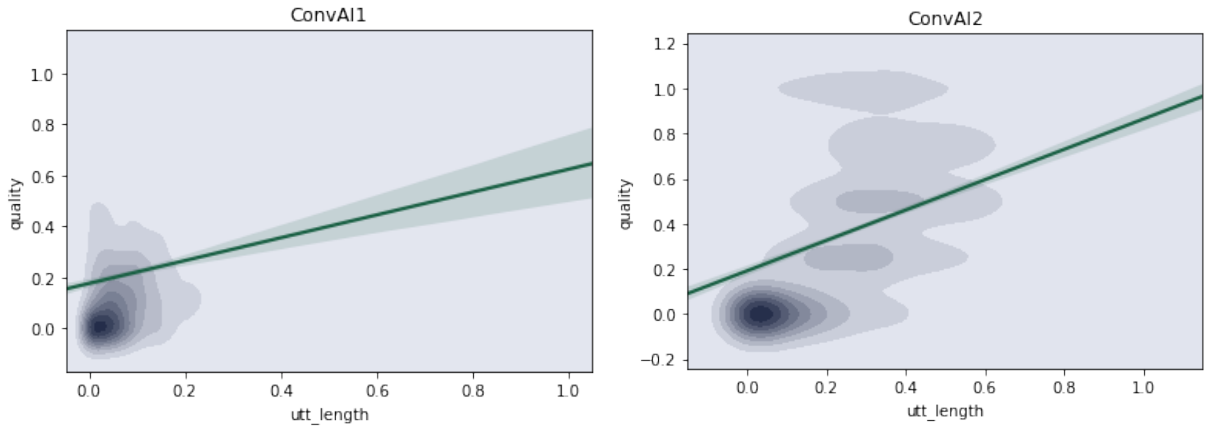
Table 3: Pearson's $r$, and Speaman's $\rho$, correlation coefficients on the two dialogue datasets' human scores and various aggreggated scores from the language models. "u-avg-d-sum" stands for averaged probabilities on utterance level and then summed up on conversation level. Most of the scores are with a confidence of $p <= 0.001$. Exceptions are GPT2-medium and GPT2 in ConvAI1 with $0.812$ and $0.212$ respectively, as well as, RUBER-U for ConvAI2, both $r$ and $\rho$, with $0.5309$ and $0.8166$, respectively.

Additionally, we investigate the effect of model size. The difference in correlation coefficients between the hyperparameter configurations is marginal and, in one of the cases, even non-existent. The most evident example is the spectrum displayed by the three GPT2 settings. Ultimately, we can conclude that smaller models perform similarly at a much smaller energy cost.

In regards to score aggregation, all the approaches unanimously show that averaging on utterance level and summing up the whole conversation is the most informative for dialogue evaluation. At the same time, the using a product or an unweighted average produce correlation coefficients very close to zero and with an extremely low significance (e.g., $p - value$ ranging from $0.4$ to $0.8$). The behavior indicates

that while utterance length is insignificant, the duration of the conversation strongly dictates its quality score.

### 4.2 Qualitative Assessment




(a) ConvAI1

(b) ConvAI2

Figure 2: Regression plots showing the relation between quality score and utterance length in the ConvAI1 and ConvAI2 datasets. The shaded area around the line represents a confidence interval.

In Figures 2a and 2b, the regression models show the interaction between the annotator quality score and the length of a conversation in ConvAI1 and ConvAI2, respectively. In both cases, the regression shows a positive trend that the longer a dialog is, the better its assessment is. We also see that in the case of ConvAI1, the confidence area is much wider than in ConvAI2. This behavior further supports the results in Table 3, where the language models have considerably lower correlation coefficients for ConvAI1.

Furthermore, we manually investigated short conversations from both datasets that also have low quality. Many of the short dialogues show that the system would indeed perform poorly by not responding at all, or the first couple of utterances would be not diverse or even the same. Thus, the annotator would terminate the session and evaluate the dialogue with a low score. In contrast, conversations that were more interactive and had longer duration also performed better in their assessment.

| **Original Context** | **Original Response** (as in dataset) | **Generated Response** (generated by transformer) |
|---|---|---|
| ”Wow! Are you man or woman?” | ”I am! i am a woman.” | ” ’m a I am a man! I” |
| ”How nice! Do you have a boyfriend?” | ”I do not. i am a single mom.” | ” ’m . . I am a virgin woman. i” |
| ”What do you mean?” | ”granted the right to accept only one religion” | ”anted, fact to be or the of” |
| ”Do you know Utrecht?” | ”granted the right to accept only one religion” | ”ind, title to use donations Dutch application” |

Table 4: Sample dialogue exchanges as originally seen in the ConvAI1 and ConvAI2 datasets together with alternative responses generated by GPT2 by just taking the most likely word. Coherent examples induces the language models to generate also good response. The top two examples have high human annotator scores, while the bottom two are rated lowly.

### 4.3 What Would a Language Model Say?

In this subsection, we report the correlation scores between the maximum probabilities for each token and the annotator scores. The intuition is that besides being renown for advancing the state-of-the-art in various NLP benchmarks, language models are prominent for being capable generators of natural language. Furthermore, Hendrycks and Gimpel (2017) have demonstrated that the maximum class probability of a neural network classifier tends to output lower values for samples that are out of distribution. Thus, we set to investigate whether the predicted maximum classes of language models can also indicate the quality of dialogues.

Although there are some studies (Wang et al., 2019) demonstrating BERT generating text, we will not consider it in this part of the work due to the nature of its masked language modeling, which does not aim at generating text. Considering GPT2 and XLNet, we look into what are the most likely words they predict for each token of the sequence instead of the original ones.

For the context of dialogue evaluation, it means that on average $max$ scores should be higher for fluent and coherent text like the one used for pretraining the language models. At the same time, erroneous samples should have lower maximum probabilities.

Firstly, we investigate the quantitative relation of the $max$ scores to human annotator scores. Similarly to what we did in Section 4.1, we have calculated the aggregated probability scores for the most likely words according to the language models:

$$lm\_dialog\_score_{max} = \sum_{u=1}^{Utterances} \left( \frac{\sum_{w=1}^{Words} P_{(w=w_{max})}}{\#Words} \right) \quad (5)$$

| Dataset | ConvAI1 | | ConvAI2 | |
|---|---|---|---|---|
| | $r$ | $\rho$ | $r$ | $\rho$ |
| **gpt2-u-avg-d-sum** | 0.133 | 0.261 | 0.193 | 0.477 |
| **gpt2-md-u-avg-d-sum** | 0.144 | 0.263 | 0.196 | 0.476 |
| **gpt2-lg-u-avg-d-sum** | 0.146 | 0.267 | 0.196 | 0.477 |
| **xlnet-u-avg-d-sum** | 0.157 | 0.263 | 0.211 | 0.471 |
| **xlnet-lg-u-avg-d-sum** | 0.137 | 0.251 | 0.209 | 0.475 |

Table 5: Pearson's correlation coefficients, $r$, and Spearman's correlation coefficients, $\rho$, on the two dialogue datasets' human scores and various aggregated scores for the $max$ word instead of the target. "u-avg-d-sum" stands for averaged probabilities on utterance level and then summed up on conversation level. All of the scores are with a confidence of $p <= 0.001$.

We present the results in Table 5. When compared to the analogous results in Table 3, we see that GPT2 and XLNet demonstrate noticeably higher correlation coefficients, especially for the dialogues in the ConvAI1 dataset. This discrepancy suggests that for some of the cases, the models can generate text that would fit better into the conversation. Since, ConvAI1 and ConvAI2 happened before the introduction of transformer-based language models, it is save to assume that the participating systems are inferior.

In Table 4, we present some short sample conversations together with a generated text by a language model. The top two examples have high scores by the human annotators, while the rest are of low quality. The model can reconstruct sensible responses that make sense and are still different from the original reply. On the other hand, whenever there is an incoherent conversation like the third and fourth examples, GPT2 and XLNet are not able to recreate a response that is either somewhat fluent or related to the current context. Another peculiarity is that the language model possesses in a sense, common knowledge. This is demonstrated by the fourth example, while in the preceding utterance, we see Utrecht, a Dutch city, and the model is then induced to predict "Dutch" as one of the response tokens.

## 5 Conclusion

In this study, we investigated whether transformer-based language models can evaluate dialogues in terms of coherency and fluency. Overall, Pearson's and Spearman's correlation coefficients demonstrate

that BERT, GPT2, and XLNet can indicate a conversation's quality without any additional supervision or reference. While, in their core, the three use the same approach, transformers, they have further structural modifications that set them apart when considered for the current problem domain.

GPT2 performs worst due to its standard language modeling approach that incorporates the least structural information about a sequence. XLNet achieves an inmprovement in terms of its correlation score by taking advantage of additional positional information when predicting a target token. Finally, BERT's next sentence prediction approach delivered the highest performance thanks to its structured approach in regards to separate utterances.

While LM-based dialogue evaluators cannot yet replace human annotators, they have additional value when compared to word-over metrics like BLUE or ones that use word-embeddings. Although they cannot completely replace human evaluators, They can support as weak indicators for quality. Additionally, we have shown that they can perform better than competing approaches like the unreferenced component of RUBER.

Furthermore, the autoregressive language models, GPT2 and XLNet, demonstrate an excellent initial aptitude for conducting dialogues. They can provide alternative responses that are also coherent with the context of a discussion.

The LM-based method in this work considers dialogues as a series paired up utterances or question-answers rather than one whole sequence. As future work, we will investigate how to extend the procedure so that it is more adept at capturing long-term information over the entire conversation.

## Acknowledgments

We acknowledge the support of the EU projects Cleopatra (GA 812997) and TAILOR (GA 952215), the Federal Ministry for Economic Affairs and Energy (BMWi) project SPEAKER (FKZ 01MK20011A), the German Federal Ministry of Education and Research (BMBF) projects and excellence clusters ML2R (FKZ 01 15 18038 A/B/C), MLwin (01S18050 D/F), ScaDS.AI (01/S18026A) as well as the Fraunhofer Zukunftsstiftung project JOSEPH.

## References

James Baker. 1975. The dragon system–an overview. *IEEE Transactions on Acoustics, speech, and signal Processing*, 23(1):24–29.

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.

Mikhail Burtsev, Varvara Logacheva, Valentin Malykh, Iulian Vlad Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, and Yoshua Bengio. 2018. The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7057–7067. Curran Associates, Inc.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of ACL 2019*, pages 2978–2988, Florence, Italy, July. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of ACL-IJNLP 2015)*, pages 445–450.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of ACL 2019*, pages 6053–6058.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of CoNLL 2018*, pages 313–323.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

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

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP 2016*, pages 2122–2132.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Varvara Logacheva, Mikhail Burtsev, Valentin Malykh, Vadim Polulyakh, and Aleksandr Seliverstov. 2018. Convai dataset of topic-oriented human-to-chatbot dialogues. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 47–57. Springer.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of ACL 2017*, pages 1116–1126.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.

Rostislav Nedelchev, Ricardo Usbeck, and Jens Lehmann. 2020. Treating dialogue quality evaluation as an anomaly detection problem. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 501–505, Marseille, France, May. European Language Resources Association.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLP 2018*, pages 2227–2237.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of EMNLP 2011*, pages 583–593. Association for Computational Linguistics.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. *arXiv preprint arXiv:1902.08832*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT 2015*, pages 196–205.

Hiroaki Sugiyama, Toyomi Meguro, and Ryuichiro Higashinaka. 2019. Automatic evaluation of chat-oriented dialogue systems using large-scale multi-references. In *Advanced Social Interaction with Agents*, pages 15–25. Springer.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Alex Wang, Kyunghyun Cho, and CIFAR Azrieli Global Scholar. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *NAACL HLT 2019*, page 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764. Curran Associates, Inc.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL 2018*, pages 2204–2213.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating commonsense in pre-trained language models. *arXiv preprint arXiv:1911.11931*.

## A Equations for Aggregating Scores on Dialogue Level

In this section, we list the different aggregation measures that we experimented with. The correlation coefficients between these aggregations scores and the human annotation are either of low values, are insignificant (low $p$-value), or both.

$$lm\_dialog\_score = \sum_{u=1}^{Utterances} \left( \sum_{w=1}^{Words} P_{(w=w)} \right) \tag{6}$$

$$lm\_dialog\_score = \frac{1}{\#Utterances} \sum_{u=1}^{Utterances} \left( \frac{\sum_{w=1}^{Words} P_{(w=w)}}{\#Words} \right) \tag{7}$$

$$lm\_dialog\_score = \frac{1}{\#Utterances} \sum_{u=1}^{Utterances} \left( \sum_{w=1}^{Words} P_{(w=w)} \right) \tag{8}$$

$$lm\_dialog\_score = \prod_{u=1}^{Utterances} \left( \sum_{w=1}^{Words} P_{(w=w)} \right) \tag{9}$$

$$lm\_dialog\_score = \prod_{u=1}^{Utterances} \left( \frac{\sum_{w=1}^{Words} P_{(w=w)}}{\#Words} \right) \tag{10}$$

$$lm\_dialog\_score = \prod_{u=1}^{Utterances} \left( \prod_{w=1}^{Words} P_{(w=w)} \right) \tag{11}$$