


## Article

# Human Annotated Dialogues Dataset for Natural Conversational Agents

Erinc Merdivan <sup>1,2,\*</sup>, Deepika Singh <sup>1,3,\*</sup>, Sten Hanke <sup>4</sup>, Johannes Kropf <sup>1</sup> ,  
Andreas Holzinger <sup>3</sup> and Matthieu Geist <sup>5</sup>

<sup>1</sup> AIT Austrian Institute of Technology, 2700 Wiener Neustadt, Austria; Johannes.Kropf@ait.ac.at

<sup>2</sup> CentraleSupélec, Université de Lorraine, CNRS, LORIA, F-57000 Metz, France

<sup>3</sup> Holzinger Group, HCI-KDD, Institute for Medical Informatics/Statistics, Medical University Graz, 8036 Graz, Austria; andreas.holzinger@medunigraz.at

<sup>4</sup> FH Joanneum Gesellschaft mbH, 8020 Graz, Austria; sten.hanke@fh-joanneum.at

<sup>5</sup> Université de Lorraine, CNRS, LIEC, F-57000 Metz, France; matthieu.geist@univ-lorraine.fr

\* Correspondence: merdivane@gmail.com (E.M.); deepika.singh@medunigraz.at (D.S.)

† These authors contributed equally to this work.

Received: 16 December 2019; Accepted: 16 January 2020; Published: 21 January 2020



**Abstract:** Conversational agents are gaining huge popularity in industrial applications such as digital assistants, chatbots, and particularly systems for natural language understanding (NLU). However, a major drawback is the unavailability of a common metric to evaluate the replies against human judgement for conversational agents. In this paper, we develop a benchmark dataset with human annotations and diverse replies that can be used to develop such metric for conversational agents. The paper introduces a high-quality human annotated movie dialogue dataset, HUMOD, that is developed from the Cornell movie dialogues dataset. This new dataset comprises 28,500 human responses from 9500 multi-turn dialogue history-reply pairs. Human responses include: (i) ratings of the dialogue reply in relevance to the dialogue history; and (ii) unique dialogue replies for each dialogue history from the users. Such unique dialogue replies enable researchers in evaluating their models against six unique human responses for each given history. Detailed analysis on how dialogues are structured and human perception on dialogue score in comparison with existing models are also presented.

**Keywords:** conversational agents; dialogue systems; chatbots

## 1. Introduction

The primary goal of intelligent dialogue systems in real-life applications is to enable efficient communication between humans and computers in a natural and coherent manner. A dialogue system requires a large amount of data to learn meaningful features and response generation strategies for building an intelligent conversational agent. Various methods, including deep learning techniques, have contributed immensely towards building dialogue systems in several real-life application domains such as natural language processing, recommender systems and question-answering systems.

It is still challenging to build a system that can understand the underlying semantics of the user input sequence, and generate coherent and meaningful responses due to limited and domain-specific dialogue datasets [1]. Previous works have developed variously structured [2,3] and unstructured [4] large datasets to train dialogue managers for the dialogue systems. One major challenge while training a dialogue manager is a lack of benchmark metrics which can be used to measure and compare performance of dialogue managers for non-task-oriented dialogue systems. Furthermore, there is no publicly available dataset with human annotations on the quality of dialogue-reply pairs to develop such metrics.

Dialogue has a certain level of abstraction due to natural language and human knowledge base; in other words, there is a very high response diversity which is unlikely to be captured by a single response [5]. Therefore, human perception of the dialogue is of utmost importance to provide diverse and unique dialogue replies depending on the dialogue history encapsulating individual's language, experience, knowledge and writing preferences. In addition, human-generated replies for the given dialogue context will help in developing robust dialogue managers, since it is not feasible to simulate good responses by using only templates. This is required in various domains but in the medical domain generally and in ambient assisted living specifically [6].

In order to address these challenges, we have developed a high-quality human annotated multi-turn movie dialogue dataset, HUMOD, from a subset of the Cornell movie dialogue dataset [7]. The collected dataset contains human annotations on fictional conversations of the movie scripts and diverse human generated replies. We have chosen Cornell movie dialogue dataset as our base dataset since it has diverse conversations on a wide span of topics, which are close to human spoken language [8], thus making the dataset ideal to train open domain dialogue systems. In [8], it is shown with quantitative and qualitative analyses that movie language is a potential source for teaching and learning spoken language features.

An example from the HUMOD dataset is shown in Figure 1. The dataset is created such that each dialogue context has two possible replies, similar to [9]—first is the actual reply of the dialogue context, which we named as positive reply and second is a reply that is sampled uniformly from the set of all possible replies, which is named as the candidate negative reply for the dialogue context. In Figure 1, a sample of the dataset is presented in three blocks, where the first block shows the 6-turn dialogue context, the second block contains a positive reply (actual reply from Cornell movie dataset) and candidate negative reply (sampled reply) and the third block shows the two human generated replies for the dialogue context. In [9,10], training is done using both positive and candidate negative replies as Next, Utterance Classification (NUC). In NUC, candidate replies can also be related to context; therefore, it is important to investigate the relevance of candidate negative reply together with actual reply. The ratings on both (positive and candidate negative) the dialogue replies from 1–5 (1: Irrelevant; 2: Weakly relevant; 3: Neutral; 4: Fairly relevant; 5: Highly relevant) are given by users. Furthermore, there are six (three for positive and three for candidate negative) human-generated replies for each dialogue context. Thus, six unique replies for each context of HUMOD dataset will help in evaluating the models against the provided human responses. The detailed descriptions of the dataset design and collection, analysis, and task formulation are discussed in Section 3. In addition, we provide benchmark performances of the supervised Hierarchical Attention Network (HAN) [11] model (with two different loss functions), Bidirectional Encoder Representations from the Transformers (BERT) [12] model and word-overlap metrics such as BLEU [13], ROUGE [14] and METEOR [15] to analyze the quality of dialogue replies for the given context. The dataset is also publicly available at: <https://github.com/erincmer/HUMOD>.

<p><b>Dialogue Context</b></p> <p><b>Speaker 1:</b> Where did you meet Miss Lawson?</p> <p><b>Speaker 2:</b> At a dinner party -- about eight months ago.</p> <p><b>Speaker 1:</b> Did you ever see her again after that?</p> <p><b>Speaker 2:</b> Yes -- several times.</p> <p><b>Speaker 1:</b> What eventually happened to your relationship with Miss Lawson?</p> <p><b>Speaker 2:</b> We stopped seeing each other.</p>
<p><b>Dialogue Replies (Humans rated 1 – 5)</b></p> <p><b>Positive:</b> Why?</p> <p><b>Negative:</b> Don't you expect me to be a little hurt?</p>
<p><b>Human Generated Replies</b></p> <p>#1: Why did you stop seeing each other?</p> <p>#2: Would you consider seeing Miss Lawson again?</p> <p>⋮</p>

**Figure 1.** A sample of 6-turn dialogue context with positive (actual) reply and candidate negative (sampled) reply and two examples of human generated replies for the dialogue context.

## 2. Related Datasets

Dialogue systems have been categorized into two groups [16]: task-oriented systems and non-task oriented systems (also known as chatbots). The aim of task-oriented systems is to assist users to complete specific tasks by understanding the inputs from the user. Manual handcrafted rules in such systems make it not only expensive and time-consuming but also limit the systems to a particular domain. Non-task oriented dialogue systems can communicate with humans on open domains, thus they can be used in real-world applications.

Dialogue systems are mainly task-oriented, designed to complete specific tasks such as airline tickets booking [17], bus information searching [18], restaurant booking [10] or railroad goods shipping [19]. Such systems perform well when the task is simple and explicit intentions of the users are well-calibrated to the system capabilities. Moreover, some of the popular domain specific datasets are bAbI simulated restaurant booking dialogues [10], Movie dialog dataset [20] and Ubuntu Dialogue Corpus dataset [4]. The Ubuntu dataset consists of large-scale unstructured dialogues with a multi-turn conversation between two persons, where negative turns of the dialogue are created by sampling. Major drawbacks of the Ubuntu dataset are that it has no annotation on context–reply pairs and no different replies for the same context.

Many recent works use conversational models for open-domain datasets such as Twitter Dialog Corpus [21] and the Chinese Weibo dataset [22] that are posts and replies from social networking sites. PERSONA-CHAT dataset [23] introduces chit-chat dialogues between crowd-sourced participants based on their given profile. The DailyDialog [24] dataset uses manual labeling with three human experts to develop a dataset with communication intention and emotion information. Moreover, the other large-scale datasets that are often used to train neural network-based conversational models are movie-subtitles datasets such as OpenSubtitles [25], Cornell Movie-Dialogue Corpus [7], Movie-DiC [26] and Movie-Triples [27]. In this work, we selected 4750 dialogues from the Cornell movie dialogue dataset as our base dialogues which are then used for human annotation as explained in the next section. A comparison of the HUMOD dataset with the existing movie dialogue datasets is shown in Table 1.

**Table 1.** A comparison of existing movie dialogue datasets with the HUMOD dataset. (\*) denotes that the HUMOD dataset can be extended by replacing the diverse replies with the original reply as explained in Figure 4.

Dataset	# of Dialogues	Human Annotated	Diverse Replies	Description
Cornell Movie-Dialogue [7]	220K	No	No	Conversation from the movie scripts.
Movie-DiC [26]	132K	No	No	American movie scripts.
Movie-Triples [27]	245K	No	No	Dialogues of three turns between two interlocutors.
OpenSubtitles [25]	36M	No	No	Movie subtitles which are not speaker-aligned.
HUMOD	28.5K *	Yes	Yes	Conversation from the movie scripts with 1 to 5 ratings and six diverse replies from humans.

In previous works of dialogue response evaluation, different unsupervised metrics have been adopted [28] to evaluate (or how not to evaluate) dialogue systems with human judgments in the Ubuntu and Twitter corpus dataset. In addition, supervised metrics [29] have been implemented that are trained to predict human judgments with or without reference reply. In this paper, we perform a correlation of human judgment with both supervised and unsupervised metrics in order to provide preliminary benchmark metrics on the HUMOD dataset.

### 3. Human Annotated Movie Dialogues Dataset (HUMOD)

#### 3.1. Dataset Design and Collection

We developed a website for data collection and used the crowdsourcing platform Amazon Mechanical Turk (AMT) (Full anonymity of the users were maintained and no ethical concerns were raised by the host institution) for collecting human annotations on selected Cornell movie dialogues. The dataset creation is performed in the following steps: First, we used simple random sampling to select 4750 dialogue contexts from the Cornell movie dataset in which each dialogue is divided into utterances. Since the dataset contains human–human dialogues, we assigned each utterance alternatively as Speaker 1 and Speaker 2 as shown in Figure 1. The dataset consists of multi-turn dialogues ranging from two to the maximum of seven turns. Then, each dialogue history is appended with an actual reply from the movie dialogue dataset and also with a candidate negative reply which is sampled uniformly from a possible set of replies, thus making the dataset size of 9500. At the beginning, basic demographics of the users were asked including Age range, Gender, and English proficiency level. These demographic variables are collected so that the dataset can be tailored as per the needs and preferences of researchers to develop conversational agents for different demographics. In the next step, each AMT user was given a task of 20 sampled dialogue conversations and was asked to rate the last reply of the dialogue conversation from 1–5, (where 1: Irrelevant reply; 2: Weakly relevant reply; 3: Neutral; 4: Fairly relevant reply; 5: Highly relevant reply) according to the dialog context and provide their relevant reply in replacement to the last reply. The total number of completed tasks were 1425.

Each dialogue–reply pair is rated by three different users, such that, for each dialog context, there are six unique responses, resulting in a total dataset size of 28,500. In order to maintain the high quality responses in the dataset, we selected responses of only those AMT users who have more than 80% hit approval rate. Each task was evaluated by a human expert and low quality tasks consisting of short or generic replies, erroneous replies consisting of gibberish text (user’s reply for given dialogue context), user’s replies in different languages (other than English) and the same ratings given to every dialogue–reply pairs were discarded.

Figures 2 and 3 present the screen shots from the website, showing how dialogue appears to the users with positive reply and candidate negative reply for the particular dialogue context. Since each task contains 20 randomly selected dialogue context-reply pairs, the chances of the same dialogue context appearing with both positive and candidate negative replies to the same user is very minimal.

## Dialogue 1

### Dialogue History:

Speaker 1: Where did you meet Miss Lawson?

Speaker 2: At a dinner party -- about eight months ago.

Speaker 1: Did you ever see her again after that?

Speaker 2: Yes -- several times.

### Candidate reply:

**Speaker 1: What eventually happend to your relationship with Miss Lawson?**

Rate the last reply (in bold) in relevance to the above conversation

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Irrelevant	Weakly Relevant	Neutral	Fairly Relevant	Highly Relevant

Provide a replacement to the candidate reply (in bold) that fits the dialogue conversation

Figure 2. Screenshot of dialogue context with positive (actual) reply.

## Dialogue 1

### Dialogue History:

Speaker 1: Where did you meet Miss Lawson?

Speaker 2: At a dinner party -- about eight months ago.

Speaker 1: Did you ever see her again after that?

Speaker 2: Yes -- several times.

### Candidate reply:

**Speaker 1: I don't know. I want it to be--**

Rate the last reply (in bold) in relevance to the above conversation

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Irrelevant	Weakly Relevant	Neutral	Fairly Relevant	Highly Relevant

Provide a replacement to the candidate reply (in bold) that fits the dialogue conversation

Figure 3. Screenshot of dialogue context with candidate negative (sampled) reply.

The six diverse replies for each dialogue context in the proposed dataset enable the extendability feature of HUMOD dataset. An example with reference to a dialogue context of how the dataset can be extended using human replies is shown in Figure 4. The seven-turn dialogue of the dataset is divided into two parts. The first part of the figure (upper block) contains the first four turns of the dialogue context and the second part (lower block) consists of all possible replies for the remaining turns of the seven-turn dialogue context. In the lower block, the second column displays the original reply from the Cornell movie dialogue dataset, and the remaining columns show the human-generated replies (four out of six) as a replacement to the original reply. Different combinations of replies in each turn can be used to augment the dataset with the caution of adding some noise. From the example in Figure 4, we analyze that, with five possible replies in each turn, a total of 125 diverse dialogues can be created, while many of the created dialogues are as coherent as the original one. This extension feature would enable to generate extra data inexpensively, which could help in training generative models to generalize better and learn to provide various replies for the same dialogue context. The extendability assumption comes from the fact that humans generated alternatives replies while being coherent with dialog context. However, it is not very beneficial to perform topic modelling or word commonality

analysis since they do not reflect coherence of alternative reply in the given dialog context and only comparing to dialogue ground reply would lead to poor results.

Dialogue Context					
{Turn 1} Speaker 1: Where did you meet Miss Lawson?					
{Turn 2} Speaker 2: At a dinner party -- about eight months ago.					
{Turn 3} Speaker 1: Did you ever see her again after that?					
{Turn 4} Speaker 2: Yes -- several times.					
#Turns	Original reply from Cornell movie dataset	Human Generated Replies			
{Turn 5} Speaker 1:	What eventually happened to your relationship with Miss Lawson?	Where did you see her?	And so, what happened?	Where and when is the last time you saw Miss Lawson?	Did anything happen during those meetings?
{Turn 6} Speaker 2:	We stopped seeing each other.	She is my best friend but I am not able to contact her since a month	Nothing came of it.	After few dates, our relationship went sore when we spoke about our career.	I lost track of her.
{Turn 7} Speaker 1:	Why?	Why did you stop?	Would you consider seeing Miss Lawson again?	Don't you expect her to be a little hurt?	You guys were so good for each other.

Figure 4. The extendability approach of the HUMOD dataset.

### 3.2. Data Analysis

This section presents the description of the HUMOD dataset and insights on human annotations. The users' demographics obtained are shown in Figure 5. The aim is to achieve human perception towards dialogue across the age groups, and we managed to achieve diversity among users, as shown in Figure 5a. The gender ratio obtained was 46% of male and 54% female, and users were also not restricted to their geography. The majority of the users (91.8%) were having advanced English level understanding as can be seen in Figure 5b, which makes the HUMOD dataset of high-quality.

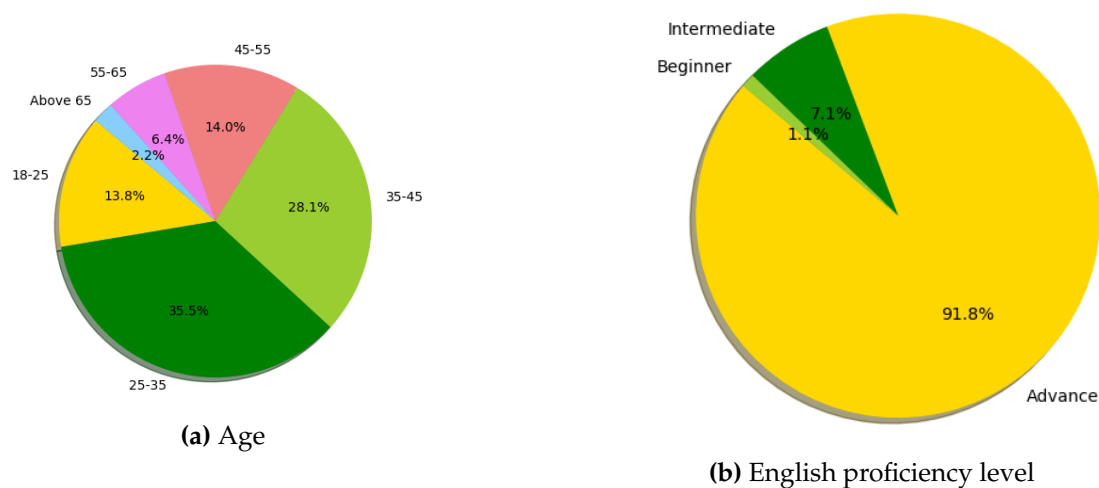


Figure 5. Users' demographics.

The number of turns per dialogue used in the dataset is given in Figure 6. Figure 7 shows the distribution of the number of words among all the obtained dialogue replies from the users. Figure 8 presents the evaluation of human generated replies on a selected HUMOD dataset. As can be seen from the figure, 2.8% of the AMT users rated Irrelevant, whereas, in the original dataset, the Irrelevant score was around 3.3% for the same dialogue–reply pairs. Figure 9 presents the histogram of user ratings for positive and negative replies. As can be seen from the figure, positive replies are rated high (4 or 5) by users, about 74% (10,610 out of 14,250) and negative replies are rated low (1 or 2), around 73% (10,404 out of 14,250). It is crucial to take into account the number of low scores in positive replies and high scores in negative replies which introduce noise in the dataset. Thus, it is important to get human judgments even for positive pairs. The negative training data can be even harmful for binary text classification [30], which may occur in the case of next utterance classification in dialogue settings. Therefore, another approach for training dialogue managers could be Positive-Unlabeled (PU) Learning [31]. In PU Learning, there is a positive set, and instead of negative there is an unlabeled dataset which can contain positive and negative samples. Existing methods [31,32] which perform well in PU Learning problems can also be applied to next utterance classification for dialogues.

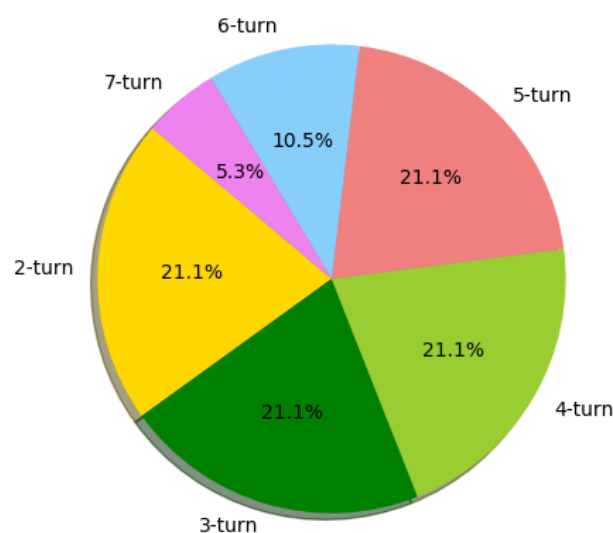


Figure 6. Number of turns in dialogue.

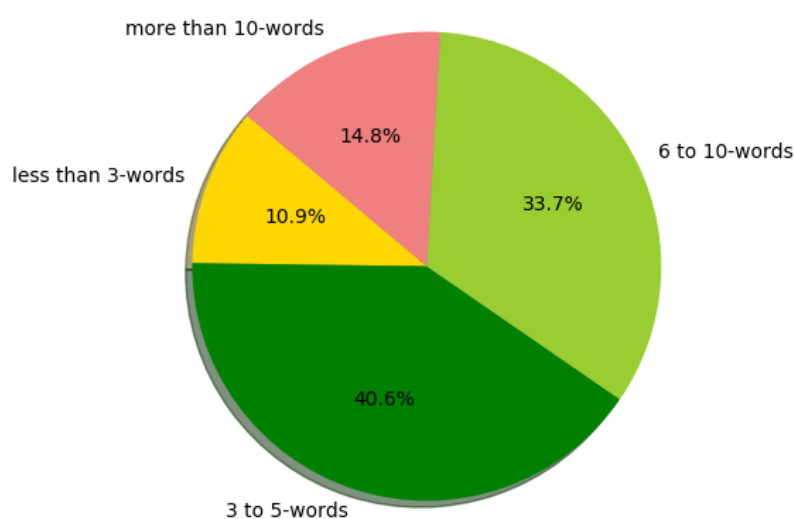


Figure 7. Number of words in human responses.

The human generated replies of HUMOD dataset are highly relevant since users were asked to provide relevant replies for the given dialogue context during data collection. However, in order to



perform initial evaluation on human generated replies, we randomly selected 30 dialogue reply pairs with six diverse replies of different turns (for example, five dialogues for each turn i.e., 2-turn, 3-turn, 4-turn, 5-turn, 6-turn, 7-turn) from the HUMOD dataset. Each diverse reply set is rated by two AMT users. The number of human generated replies for evaluation could be higher with increasing cost, but a strong pattern of relevancy can be easily seen in these sampled human generated replies.

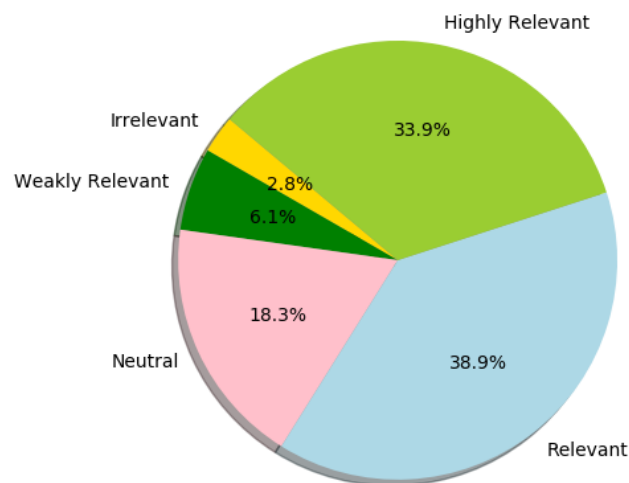


Figure 8. Evaluation of human responses on the selected HUMOD dataset.

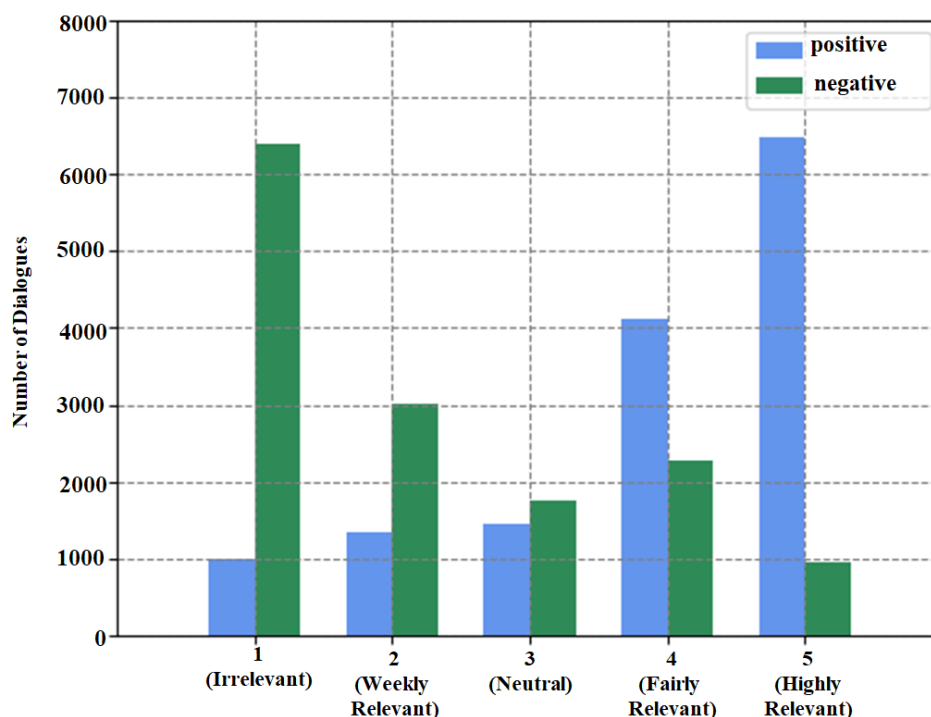


Figure 9. Human scores vs. Positive and Candidate Negative Dialogue Pairs.

To evaluate human agreement on their responses with each other, we calculated weighted Cohen's kappa score [33] between human ratings. We calculated weighted kappa score for different configurations of three ratings for each dialogue context and reply pair. From the three ratings of each reply, we calculated weighted kappa score for the closest two (as a majority voting) ratings, the highest two ratings, the lowest two ratings, and on the random selection of two ratings. For example, if a dialogue context-reply pair is rated (5, 4 and 1), we keep the closest two (5 and 4) and randomly assign them to Rater 1 and 2. Results for all different configurations can be seen in Table 2.



**Table 2.** Inter-annotator agreement.

Rater	Cohen's Kappa Score ( $\kappa$ )
Closest two ratings	0.86
Lowest two ratings	0.57
Highest two ratings	0.55
Random two ratings	0.42

### 3.3. Task Formulation

We provide a benchmark performance with supervised models and word-overlap metrics on the proposed HUMOD dataset. The task formulated is to predict human ratings for given dialogue history–reply pair from Irrelevant (1) to Highly-relevant (5). The dataset will enable researchers to test their dialogue reply metrics against human ratings.

## 4. Methods

This section provides details about the supervised models and word-overlap metrics, which are used to predict the human scores for dialogue replies. For supervised models, a network inspired from hierarchical attention networks (HAN) is designed as shown in Figure 10. Recurrent Neural Network (RNN) based models are implemented in a way such that it does not require a reference answer, since it makes them more applicable but difficult to train in comparison to the models which uses reference response. Our model is the same as the Automatic Dialogue Evaluation Model (ADEM) [29] model with an additional attention layer. In ADEM, training and evaluation was performed with and without reference reply. Although ADEM results without reference reply is lower in comparison to with the reference reply, we preferred to train without a reference as it is not practical to have reference reply in real life applications. In addition, we used a Bidirectional Encoder Representations from Transformers (BERT) supervised model for comparison with HAN and word-overlap metrics. Word-overlap metrics require reference answers unlike supervised models and only use reference reply without any context information, which makes them perform poorly. These metrics are easy and inexpensive to run without the need for training and can be applied as long as there is a reference text. Supervised models require training to predict the dialogue reply score which makes it difficult to generalize and expensive to use in different domains and tasks. We used metrics such as BLEU [13], METEOR [15] and ROUGE [14], where the first two are used widely in translation and the last one for assessing the quality of the summarized text.

### 4.1. BLEU

BLEU [13] score calculates the precision of n-grams of machine-generated dialogue replies in human replies. Often, the brevity penalty is used to avoid short sentences. From different n-gram choices, the most common one is  $n = 4$  where the weighted average of different BLEU (1 to 4) scores are used to evaluate the machine-generated replies. BLEU first calculates a modified precision score for each n-gram length as below:

$$P_n(r, \hat{r}) = \frac{\sum_{ngr} Count_{matched}(ngr)}{\sum_k Count(ngr)}, \quad (1)$$

where  $ngr$  represents all possible n-grams of length  $n$  in hypothesis sentences. Later,

$$BLEU-4 = b(r, \hat{r}) \exp \left( \sum_{n=1}^4 0.25 \log P_n(r, \hat{r}) \right). \quad (2)$$

To avoid shorter sentences, the modified precision score is often multiplied with a brevity penalty to achieve the final score.

#### 4.2. ROUGE

The ROUGE [14] score originally calculates the recall of n-grams of human dialogue replies in machine-generated dialogue replies. ROUGE can be extended to ROUGE-L, which is a combination of recall and precision ROUGE scores based on longest matching sequence (LCS). LCS only requires sentence level word order matches and it allows other words to appear between words of LCS. LCS does not need n-gram to be predefined:

$$R = \max_j \frac{l(r, \hat{r}_{i_j})}{|\hat{r}_{i_j}|}, \quad (3)$$

$$P = \max_j \frac{l(r_i, \hat{r}_{i_j})}{|r_i|}, \quad (4)$$

$$\text{ROUGE} = \frac{(1 + \beta^2)RP}{R + \beta^2 P}. \quad (5)$$

#### 4.3. METEOR

METEOR [15] score aligns human generated reply and machine-generated reply. This alignment is done word by word with an order of exact match, Porter stemming match or WordNet synonym match. It computes the parametric harmonic mean ( $F_{mean}$ ) between unigram recall and unigram precision:

$$P = \frac{m}{t}, \quad (6)$$

$$R = \frac{m}{r}, \quad (7)$$

$$F_{mean} = \frac{P \cdot R}{\alpha P + (1 - \alpha) R}, \quad (8)$$

$$Pen = \gamma \left( \frac{ch}{m} \right)^\theta, \quad (9)$$

$$\text{METEOR} = (1 - Pen) F_{mean}, \quad (10)$$

where  $t$  and  $r$  are the total numbers of unigrams in the translation and the reference.  $m$  represents the number of mapped unigrams between reference and hypothesis sentences. Penalty term ( $Pen$ ) is computed so that it takes into account matched unigrams between hypothesis and reference that are in the same order.

#### 4.4. Hierarchical Attention Network-Based Models

In this work, we implemented two separate networks, one for dialogue context and another for the dialogue reply. Context encoder is a Hierarchical Attention Network (HAN) [11] as shown in Figure 10. It encodes dialogue context to a dialog context vector  $c$ . Reply encoder network is a biLSTM [34] network with attention mechanism on top to convert reply to a reply vector  $\hat{r}$  (Figure 11).

We implemented two different loss functions after the dialogue context and reply are encoded to vector representation. The first loss is the cross-entropy loss which classifies concatenated vectors of dialogue context and reply into five classes (1–5):

$$\text{HAN-R(MSE)}(c, \hat{r}) = \sum_i [FC([c_i, \hat{r}_i]) - h_i]^2, \quad (11)$$

where  $h$  represents average human scores,  $c$  represents the context and  $\hat{r}$  is reply vector.  $FC$  is the fully connected layer which outputs the score for the given context and reply.

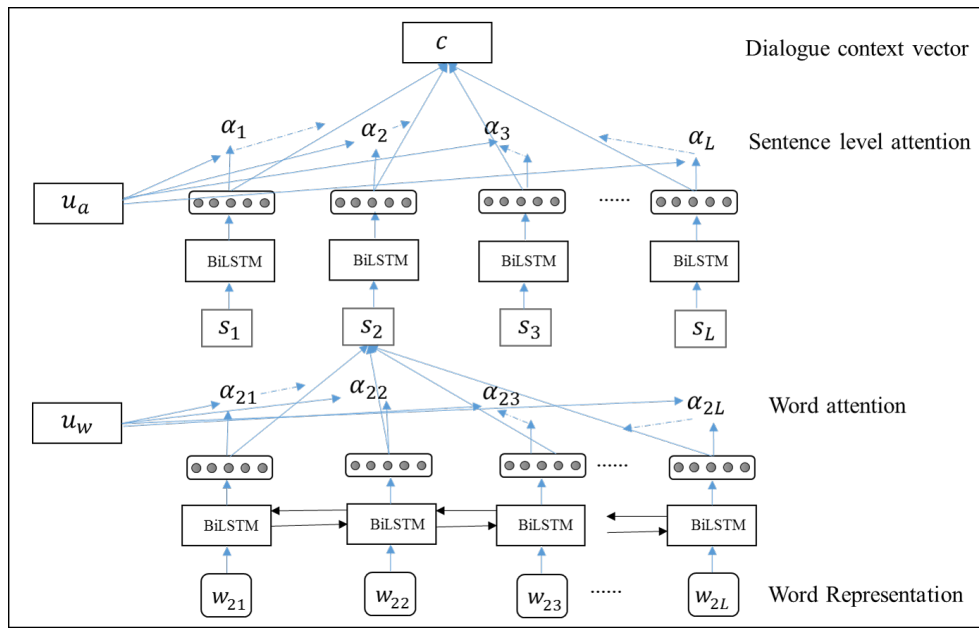


Figure 10. Hierarchical attention network for the dialogue context.

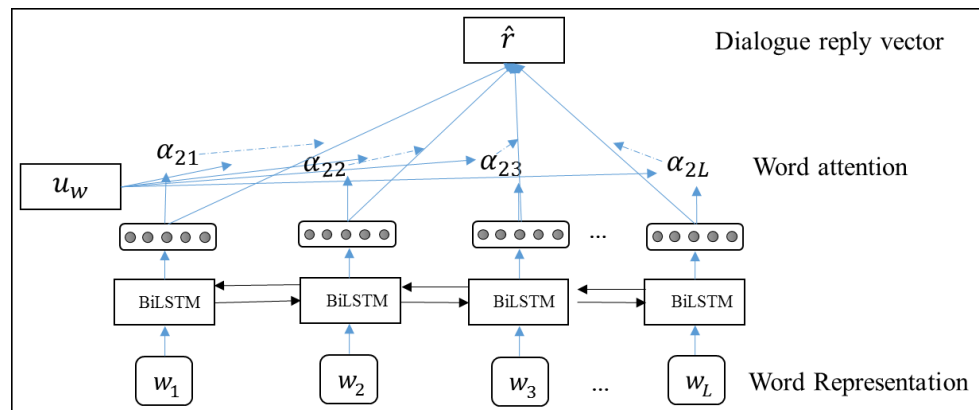


Figure 11. Dialogue reply encoder.

Second loss, as in Equation (11), is trained with mean squared error against human score and uses a concatenated vector. We did not use cosine similarity between dialogue context and reply since the authors in [35] showed that ADEM, which is similar to ours, creates response embedding with very low vector spread in the embedding space. In dialogue, there are many alternative replies for the same context, and the same reply can be a good fit to very different contexts and common replies that occur a lot and fit to a lot of different contexts. Due to these observations, when a dialogue manager is trained using cosines distance and make embedding of context and reply similar, eventually very different texts become similar and may collapse to very small region in embedding space.

#### 4.5. Bidirectional Encoder Representations from Transformers (BERT)

BERT [12] is a bidirectional language model that allows models to learn both left and right context in all layers. BERT is pretrained with two novel methods that are “Masked Language Model (MLM)” and “Next sentence prediction” and uses Transformer [36] with attention instead of recurrent networks like LSTM. The Next sentence prediction method is more related to our task since it trains BERT to learn the relationship between sentences, which may improve performance in case of dialogue context and reply scoring. The BERT model has achieved state-of-the-art performance on various Natural language processing (NLP) tasks and also outperformed human performance in question-answering

tasks. Therefore, we fine-tuned BERT (from tensorflow-hub) on the HUMOD dataset as shown in Figure 12 in order to compare performance with other approaches. As can be seen from Figure 12, BERT takes input as tokenized dialogue context and dialogue reply separated with SEP token and outputs a final vector representation of dialogue context and reply pair. Later, this vector is classified to a relevance score of 1–5 with additional dense layer on top of BERT.

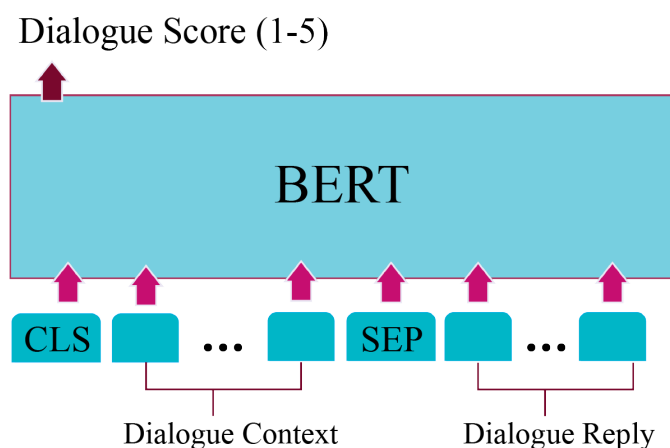


Figure 12. BERT used as a dialogue reply scorer.

## 5. Experiments

We performed a comparison of supervised and word-overlap metrics to see how they are correlated with human judgments. The dataset is divided into 8500 context-reply pairs for the train set and 1000 context-reply pairs for the test set.

No dialogue context is shared between train and test set. Since for each dialogue we have scores of three judges, we took the average score of three judges. Both supervised and word-overlap approaches are evaluated on the same test set. For word-overlap metrics, we normalized average human scores into the 0–1 range and calculated metrics using the NLG Eval toolkit [37]. Both the context encoder and the reply encoder use 50-dimensional GloVe word vectors [38], and the dimension of 100 was chosen for each biLSTM hidden state for the HAN-based model. For the model which uses BERT, input is constructed as dialogue context and dialogue reply separated with a special token, and the final encoded vector is classified into one of the five classes.

## 6. Empirical Results

The preliminary results of HUMOD dataset with existing supervised and word-overlap metrics is shown in Table 3.

Table 3. Correlation of different models with human scores.

Models	Correlation ( <i>p</i> -Value)
BLEU-4	0.055 (0.08)
ROUGE	−0.035 (0.26)
METEOR	−0.017 (0.59)
HAN-R(CE)	0.138 (<0.001)
HAN-R(MSE)	0.128 (0.003)
BERT	0.602 (<0.001)

We provided the benchmark results for the overall correlation of human judgment with different models. Although supervised models are correlated up to some degree, it is still far from applicable to use in dialogue reply scoring as widely as translation scores are used to evaluate machine translation models both in terms of human correlation and ease of use.

As can be seen from Table 3, the BERT model outperformed the word-overlap metrics and HAN model. Since the BERT model takes advantage of the language model and can be fine-tuned according to other dataset, in this experiment, we used BERT with pretrained weights and fine-tuned it for the HUMOD dataset, which may explain the performance difference of supervised models. In addition, correlation of BERT with human judges is performed to investigate the behavior of the network against different dialogue turns (shown in Table 4). We tested BERT performance on different turn dialogues to see how context length affects the performance of dialogue measure. 2-turn dialogues correlation is found to be lowest, which can be due to hardness to evaluate the dialogue reply score for very short dialogues since it contains very little context, which may increase humans' own judgement and bias. Similarly, for long-dialogue conversation, it is slightly harder for the system to contain the context and make a good understanding of the fitness of dialogue context and reply pairs.

**Table 4.** Correlation of BERT against different turns with human scores.

Dialogue Turn	Correlation ( <i>p</i> -Value)
2-turn	0.52 (<0.001)
3-turn	0.58 (<0.001)
4-turn	0.67 (<0.001)
5-turn	0.61 (<0.001)
6-turn	0.66 (<0.001)
7-turn	0.59 (<0.001)

## 7. Conclusions and Future Work

This paper presents the human annotated movie dialogue dataset (HUMOD) for research to develop benchmark metrics for comparison of different models on human scores and generated replies. The detailed description of the dataset construction and statistics is provided. The availability of the HUMOD dataset opens up various possibilities for research and development of complex dialogue systems for real-life applications. Different replies for the same context as well as dialogue ratings can be used to develop a metric to compare methods such as BLEU. Another interesting usage of a unique diverse reply is to train generative models that generate diverse dialogue replies which may make dialogue managers more human-like in real-life applications. We present the preliminary results to provide baselines with supervised and word-overlap metrics. HAN provides better results in comparison to word-overlap metrics since these approaches do not use any context. The BERT model outperforms the HAN model and provides good correlation on the human dialogue score; however, it is harder to train and requires fine-tuning for each different dataset.

In future work, we will work on a metric that uses dialogue contexts and replies to produce a robust score without any training. We will also focus on generative models that can leverage all possible replies for a given context rather than a single context–reply pair. Another interesting approach is to generate diverse replies from single context–reply pairs and to use such created data as augmentation for dialogue managers. All this is relevant for the upcoming research stream of explainable AI, where NLU [39] plays a particular role, e.g., in the generation of human-understandable explanations [40,41], where it is very useful for the development of future human-AI interfaces.

**Author Contributions:** Conceptualization, E.M. and D.S.; methodology, E.M. and D.S.; software, E.M. and D.S.; validation, E.M. and D.S.; formal analysis, E.M. and D.S.; investigation, E.M. and D.S.; resources, E.M. and D.S.; data curation, E.M. and D.S.; writing—original draft preparation, E.M. and D.S.; writing—review and editing, E.M., D.S., A.H. and M.G.; visualization, E.M. and D.S.; supervision, S.H., J.K., A.H. and M.G.; project administration, S.H. and J.K.; funding acquisition, S.H., J.K. and A.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work has been funded by the European Union Horizon2020 MSCA ITN ACROSSING project (GA no. 616757).

**Acknowledgments:** The authors would like to thank the members of the ACROSSING project's consortium for their valuable inputs. We are also thankful to all the participants of the dialogue dataset survey.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

HUMOD	Human annotated movie dialogue dataset
NUC	Next utterance classification
AMT	Amazon mechanical turk
HAN	Hierarchical attention network
BERT	Bidirectional encoder representations from transformers
NLP	Natural language processing
LSTM	Long short term memory
LCS	Longest matching sequence
PU	Positive-unlabeled
CE	Context encoder
MSE	Mean squared error

## References

- Petukhova, V.; Gropp, M.; Klakow, D.; Schmidt, A.; Eigner, G.; Topf, M.; Srb, S.; Motlicek, P.; Potard, B.; Dines, J.; et al. *The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues*; Technical Report; European Language Resources Association (ELRA): Paris, France, 2014.
- Henderson, M.; Thomson, B.; Williams, J.D. The second dialog state tracking challenge. In *Proceedings of the SIGDIAL 2014 Conference*, Philadelphia, PA, USA, 18–20 June 2014.
- Williams, J.; Raux, A.; Ramachandran, D.; Black, A. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, Metz, France, 22–24 August 2013.
- Lowe, R.; Pow, N.; Serban, I.; Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the SIGDIAL 2015 Conference*, Prague, Czech Republic, 2–4 September 2015.
- Artstein, R.; Gandhe, S.; Gerten, J.; Leuski, A.; Traum, D. Semi-formal evaluation of conversational characters. In *Languages: From formal to natural*; Springer: Berlin/Heidelberg, Germany, 2009.
- Holzinger, A. User-Centered Interface Design for disabled and elderly people: First experiences with designing a patient communication system (PACOSY). In *Computer Helping People with Special Needs, ICCHP 2002, Lecture Notes in Computer Science (LNCS 2398)*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 34–41. [[CrossRef](#)]
- Danescu-Niculescu-Mizil, C.; Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, Portland, OR, USA, 23 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011.
- Forchini, P. Spontaneity reloaded: American face-to-face and movie conversation compared. In *Proceedings of the Corpus Linguistics Conference 2009 (CL2009)*, Liverpool, UK, 20–23 July 2009.
- Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. On the evaluation of dialogue systems with next utterance classification. In *Proceedings of the SIGDIAL 2016 Conference*, Los Angeles, CA, USA, 13–15 September 2016.
- Bordes, A.; Boureau, Y.L.; Weston, J. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, 24–26 April 2017.
- Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, San Diego, CA, USA, 12–17 June 2016.
- Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2–7 June 2019; Volume 1. (Long and Short Papers).



13. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association For Computational Linguistics (ACL), Philadelphia, PA, USA, 7–12 July 2002.
14. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004.
15. Banerjee, S.; Lavie, A. 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, Ann Arbor, MI, USA, 9 June 2005.
16. Chen, H.; Liu, X.; Yin, D.; Tang, J. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 2. [\[CrossRef\]](#)
17. Zue, V.; Seneff, S.; Polifroni, J.; Phillips, M.; Pao, C.; Goodine, D.; Goddeau, D.; Glass, J. PEGASUS: A spoken dialogue interface for online air travel planning. *Speech Commun.* **1994**, *15*, 331–340. [\[CrossRef\]](#)
18. Raux, A.; Langner, B.; Bohus, D.; Black, A.W.; Eskenazi, M. Let's Go Public! Taking a spoken dialog system to the real world. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
19. Allen, J.F.; Miller, B.W.; Ringger, E.K.; Sikorski, T. A robust system for natural spoken dialogue. In Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL), Santa Cruz, CA, USA, 23–28 June 1996.
20. Dodge, J.; Gane, A.; Zhang, X.; Bordes, A.; Chopra, S.; Miller, A.; Szlam, A.; Weston, J. Evaluating prerequisite qualities for learning end-to-end dialog systems. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
21. Ritter, A.; Cherry, C.; Dolan, W.B. Data-driven response generation in social media. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, UK, 27–31 July 2011.
22. Wang, H.; Lu, Z.; Li, H.; Chen, E. A dataset for research on short-text conversations. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), Washington, DC, USA, 18–21 October 2013.
23. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv* **2018**, arXiv:1801.07243.
24. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Taipei, Taiwan, 27 November–1 December 2017.
25. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 21–22 May 2012.
26. Banchs, R.E. Movie-DiC: A movie dialogue corpus for research and development. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), Jeju Island, Korea, 8–14 July 2012.
27. Serban, I.V.; Sordoni, A.; Bengio, Y.; Courville, A.C.; Pineau, J. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI, Phoenix, AZ, USA, 12–17 February 2016.
28. Liu, C.W.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv* **2016**, arXiv:1603.08023.
29. Lowe, R.; Noseworthy, M.; Serban, I.V.; Angelard-Gontier, N.; Bengio, Y.; Pineau, J. Towards an automatic Turing test: Learning to evaluate dialogue responses. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver, BC, Canada, 30 July–4 August 2017.
30. Li, X.L.; Liu, B.; Ng, S.K. Negative training data can be harmful to text classification. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), Cambridge, MA, USA, 9–11 October 2010.
31. Li, X.L.; Liu, B. Learning from positive and unlabeled examples with different data distributions. In Proceedings of the European Conference on Machine Learning (ECML). Springer, Porto, Portugal, 3–7 October 2005.
32. Merdivan, E.; Loghmani, M.R.; Geist, M. Reconstruct & Crush Network. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.



33. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 213. [[CrossRef](#)] [[PubMed](#)]
34. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [[CrossRef](#)] [[PubMed](#)]
35. Sai, A.B.; Gupta, M.D.; Khapra, M.M.; Srinivasan, M. Re-evaluating ADEM: A Deeper Look at Scoring Dialogue Responses. *AAAI* **2019**, *30*, 1. [[CrossRef](#)]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
37. Sharma, S.; El Asri, L.; Schulz, H.; Zumer, J. Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation. *arXiv* **2017**, arXiv:1706.09799.
38. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014.
39. Holzinger, A.; Kieseberg, P.; Weippl, E.; Tjoa, A.M. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. In *Springer Lecture Notes in Computer Science LNCS 11015*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1–8. [[CrossRef](#)]
40. Hudec, M.; Bednárová, E.; Holzinger, A. Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language. *J. Off. Stat.* **2018**, *34*, 981. [[CrossRef](#)]
41. Holzinger, A.; Kickmeier-Rust, M.; Mueller, H. KANDINSKY Patterns as IQ-Test for machine learning. In *Springer Lecture Notes LNCS 11713*; Springer Nature: Cham, Switzerland, 2019; pp. 1–14. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).