Game-Based Video-Context Dialogue

Ramakanth Pasunuru and **Mohit Bansal**
UNC Chapel Hill
{ram, mbansal}@cs.unc.edu

Abstract

Current dialogue systems focus more on textual and speech context knowledge and are usually based on two speakers. Some recent work has investigated static image-based dialogue. However, several real-world human interactions also involve dynamic visual context (similar to videos) as well as dialogue exchanges among multiple speakers. To move closer towards such multimodal conversational skills and visually-situated applications, we introduce a new video-context, many-speaker dialogue dataset based on live-broadcast soccer game videos and chats from Twitch.tv. This challenging testbed allows us to develop visually-grounded dialogue models that should generate relevant temporal and spatial event language from the live video, while also being relevant to the chat history. For strong baselines, we also present several discriminative and generative models, e.g., based on tridirectional attention flow (TriDAF). We evaluate these models via retrieval ranking-recall, automatic phrase-matching metrics, as well as human evaluation studies. We also present dataset analyses, model ablations, and visualizations to understand the contribution of different modalities and model components.

1 Introduction

Dialogue systems or conversational agents which are able to hold natural, relevant, and coherent interactions with humans have been a long-standing goal of artificial intelligence and machine learning. There has been a lot of important previous work in this field for decades (Weizenbaum, 1966; Isbell et al., 2000; Rambow et al., 2001; Rieser et al., 2005; Georgila et al., 2006; Rieser and Lemon, 2008; Ritter et al., 2011), including recent work on introduction of large textual-dialogue datasets (e.g., Lowe et al. (2015); Serban et al. (2016)) and end-to-end neural network based models (Sordoni et al., 2015; Vinyals and Le, 2015; Su et al., 2016; Luan et al., 2016; Li et al., 2016; Serban et al., 2017a,b).

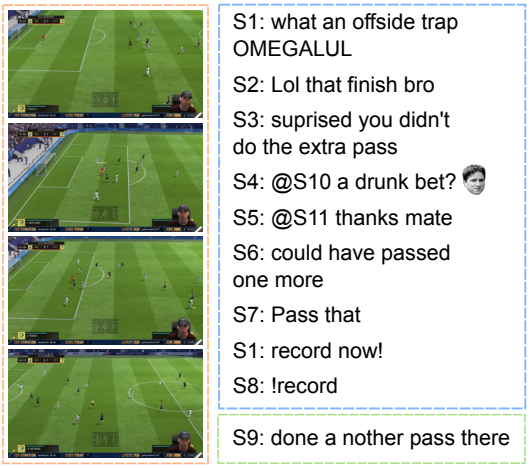

Figure 1: Sample example from our many-speaker, video-context dialogue dataset, based on live soccer game chat. The task is to predict the response (bottom-right) using the video context (left) and the chat context (top-right).

Current dialogue tasks are usually focused on the textual or verbal context (conversation history). In terms of multimodal dialogue, speech-based spoken dialogue systems have been widely explored (Eckert et al., 1997; Singh et al., 2000; Young, 2000; Janin et al., 2003; Celikyilmaz et al., 2017; Wen et al., 2015; Su et al., 2016; Mrkšić et al., 2016), as well as work on gesture and haptics based dialogue (Johnston et al., 2002; Cassell, 1999; Foster et al., 2008). In order to address the additional advantage of using visually-grounded context knowledge in dialogue, recent work introduced the visual dialogue task (Das et al., 2017; de Vries et al., 2017; Mostafazadeh et al., 2017). However, the visual context in these tasks is lim-

We release all data, code, and models at: https://github.com/ramakanth-pasunuru/video-dialogue

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 125–136
Brussels, Belgium, October 31 - November 4, 2018.

ited to one static image. Moreover, the interactions are between two speakers with fixed roles (one asks questions and the other answers).

Several situations of real-world dialogue among humans involve more 'dynamic' visual context, i.e., video-style information of the world moving around us (both spatially and temporally). Further, several human conversations involve more than two speakers, with changing roles. In order to develop such dynamically-visual multimodal dialogue models, we introduce a new 'many-speaker, video-context chat' testbed, along with a new dataset and models for the same. Our dataset is based on live-broadcast soccer (FIFA-18) game videos from the 'Twitch.tv' live video streaming platform, along with the spontaneous, many-speaker live chats about the game. This challenging testbed allows us to develop dialogue models where the generated response is required to be relevant to the temporal and spatial events in the live video, as well as be relevant to the chat history (with potential impact towards video-grounded applications such as personal assistants, intelligent tutors, and human-robot collaboration).

We also present several strong discriminative and generative baselines that learn to retrieve and generate bimodal-relevant responses. We first present a triple-encoder discriminative model to encode the video, chat history, and response, and then classify the relevance label of the response. We then improve over this model via tridirectional attention flow (TriDAF). For the generative models, we model bidirectional attention flow between the video and textual chat context encoders, which then decodes the response. We evaluate these models via retrieval ranking-recall, phrase-matching metrics, as well as human evaluation studies. We also present dataset analysis as well as model ablations and attention visualizations to understand the contribution of the video vs. chat modalities and the model components.

2 Related Work

Early dialogue systems had components of natural language (NL) understanding unit, dialogue manager, and NL generation unit (Bates, 1995). Statistical learning methods were used for automatic feature extraction (Dowding et al., 1993; Mikolov et al., 2013), dialogue managers incorporated reward-driven reinforcement learning (Young et al., 2013; Shah et al., 2016), and the generation units have been extended with seq2seq neural network models (Vinyals and Le, 2015; Serban et al., 2016; Luan et al., 2016).

In addition to the focus on textual dialogue context, using multimodal context brings more potential for having real-world grounded conversations. For example, spoken dialogue systems have been widely explored (Singh et al., 2000; Gurevych and Strube, 2004; Georgila et al., 2006; Eckert et al., 1997; Young, 2000; Janin et al., 2003; De Mori, 2007; Wen et al., 2015; Su et al., 2016; Mrkšić et al., 2016; Hori et al., 2016; Celikyilmaz et al., 2015, 2017), as well as gesture and haptics based dialogue (Johnston et al., 2002; Cassell, 1999; Foster et al., 2008). Additionally, dialogue systems for digital personal assistants are also well explored (Myers et al., 2007; Sarikaya et al., 2016; Damacharla et al., 2018). In the visual modality direction, some important recent attempts have been made to use static image based context in dialogue systems (Das et al., 2017; de Vries et al., 2017; Mostafazadeh et al., 2017), who proposed the 'visual dialog' task, where the human can ask questions on a static image, and an agent interacts by answering these questions based on the previous chat context and the image's visual features. Also, Celikyilmaz et al. (2014) used visual display information for on-screen item resolution in utterances for improving personal digital assistants.

In contrast, we propose to employ dynamic video-based information as visual context knowledge in dialogue models, so as to move towards video-grounded intelligent assistant applications. In the video+language direction, previous work has looked at video captioning (Venugopalan et al., 2015) as well as Q&A and fill-in-the-blank tasks on videos (Tapaswi et al., 2016; Jang et al., 2017; Maharaj et al., 2017) and interactive 3D environments (Das et al., 2018; Yan et al., 2018; Gordon et al., 2017; Anderson et al., 2017). There has also been early related work on generating sportscast commentaries from simulation (RoboCup) soccer videos represented as non-visual state information (Chen and Mooney, 2008). Also, Liu et al. (2016a) presented some initial ideas on robots learning grounded task representations by watching and interacting with humans performing the task (i.e., by converting human demonstration videos to Causal And-Or graphs). On the other hand, we propose a new video-chat dataset where the

Figure 2: Sample page of live broadcast of FIFA-18 game on twitch.tv with concurrent user chat.

dialogue models need to generate the next response in the sequence of chats, conditioned both on the raw video features as well as the previous textual chat history. Moreover, our new dataset presents a many-speaker conversation setting, similar to previous work on meeting understanding and Computer Supported Cooperative Work (CSCW) (Janin et al., 2003; Waibel et al., 2001; Schmidt and Bannon, 1992). In the live video stream direction, Fu et al. (2017) and Ping and Chen (2017) used real-time comments to predict the frame highlights in a video, and Barbieri et al. (2017) presented emotes and troll prediction.

3 Twitch-FIFA Dataset

3.1 Dataset Collection and Processing

For our new video-context dialogue task, we used the publicly accessible Twitch.tv live broadcast platform, and collected videos of soccer (FIFA-18) games along with the users' live chat conversations about the game. This dataset has videos involving various realistic human actions and events in a complex sports environment and hence serves as a good testbed and first step towards multimodal video-based dialogue data. An example is shown in Fig. 1 (and an original screenshot example in Fig. 2), where the users perform a complex 'many-speaker', 'multimodal' dialogue. Overall, we collected 49 FIFA-18 game videos along with their users' chat, and divided them into 33 videos for training, 8 videos for validation, and 8 videos for testing. Each such video is several hours long, providing a good amount of data (Table 2).

To extract triples (instances) of video context, chat context, and response from this data, we divide these videos based on the fixed time frames instead of fixed number of utterances in order to maintain conversation topic clusters (because of the sparse nature of chat utterances count over the time). First, we use 20-sec context windows to extract the video clips and users utterances in this time frame, and use it as our video and chat contexts, resp. Next, the chat utterances in the immediately-following 10-sec window (response window) that do not overlap with the next instance's context window are considered as potential responses.[1] Hence, there are only two instances (triples) in a 60-sec long video, i.e., 20-sec video+chat context window and 10-sec response window, and there is no overlap between the instances. Now, out of these potential responses, to only allow the response that has at least some good coherence and relevance with the chat context's topic, we choose the first (earliest) response that has high similarity with some other utterance in this response window (using 0.5 BLEU-4 threshold, based on manual inspection).[2]

	Relevance to Video+Chat
filtered response wins	34%
1st response wins	3%
Non-distinguishable	63% (56 both-good, 7 both-bad)

Table 1: Human evaluation of our dataset, comparing our filtered responses versus the first response in the window (for relevance w.r.t. video and chat contexts).

Human Quality Evaluation of Data Filtering Process: To evaluate the quality of the responses that result from our filtering process described above, we performed an anonymous (randomly shuffled w/o identity) human comparison between the response selected by our filtering process vs. the first response from the response window without any filtering, based on relevance w.r.t. video and chat context. Table 1 presents the results on 100 sample size, showing that humans in a blind-test found 90% (34+56) of our filtered responses as valid responses, verifying that our response selection procedure is reasonable. Furthermore, out of these 90% valid responses, we found that 55% are chat-only relevant, 11% are video-only relevant, and 24% are both video+chat relevant.

In order to make the above procedure safe and to make the dataset more challenging, we also discourage frequent responses (top-20 most-frequent

[1] We use non-overlapping windows because: (1) the utterances are non-uniformly distributed in time and hence if we have a shifting window, sometimes a particular data instance/chunk becomes very sparse and contains almost zero utterances; (2) we do not want overlap between response of one window with the context of the next window, so as to avoid the encoder already having seen the response (as part of context) that the decoder needs to generate for the other window.

[2] Based on intuition that if multiple speakers are saying the same response in that 10-second window, then this response should be more meaningful/relevant w.r.t. chat context.

Statistics	Train	Val	Test
#Videos	33	8	8
Total Hours	58.4	11.9	15.4
Final Filtered #Instances	10,150	2,153	2,780
Avg. Chat Context Length	69.0	63.5	71.2
Avg. Response Length	6.5	6.5	6.1

Table 2: Twitch-FIFA dataset's chat statistics (lengths are defined in terms of number of words).

generic utterances) unless no other response satisfies the similarity condition, hence suppressing the frequent responses.[3] If we couldn't find any utterance based on the multi-response matching procedure described above, then we just consider the first utterance in the 10-second window as the response.[4] We also make sure that the chat context window has at least 4 utterances, otherwise we exclude that context window and also the corresponding response window from the dataset. After all this processing, our final resulting dataset contains $10,510$ samples in training, $2,153$ samples in validation, and $2,780$ samples in test.[5]

3.2 Dataset Analysis

Dataset Statistics Table 2 presents the full statistics on train, validation, and test sets of our Twitch-FIFA dataset, after the filtering process described in Sec. 3.1. As shown, the average chat context length in the dataset is around 68 words, and the average response length is 6.3 words.

Chat Context Size Fig. 3 presents the study of number of utterances in the chat context vs. the number of such training samples. As we limit the minimum number of utterances to 4, chat context with less than 4 utterances is not present in the dataset. From the Fig. 3, it is clear that as the number of utterances in the chat context increases, the number of such training samples decrease.

Frequent Words Fig. 4 presents the top-20 frequent words (excluding stop words) and their corresponding frequency in our Twitch-FIFA dataset. Most of these frequent words are related to soccer vocabulary. Also, some of these frequent words are twitch emotes (e.g. 'kappa', 'inceptionlove').

[3]Note that this filtering suppresses the performance of simple frequent-response baseline described in Sec. 4.1.

[4]Other preprocessing steps include: omit the utterances in the response window which refer to a speaker name out of the current chat context; remove non-representative utterances, e.g., those with hyperlinks; replace (anonymize) all the user identities mentioned in the utterances with a common tag (i.e., anonymizing due to similar intuitions from the Q&A community (Hermann et al., 2015)).

[5]Note that this is substantially larger than or comparable to most current video captioning datasets. We plan to further extend our dataset based on diverse games and video types.

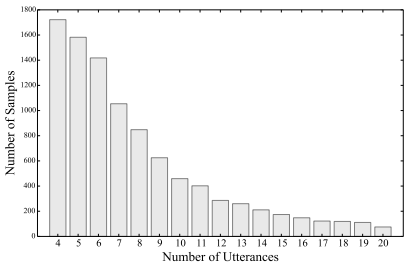

Figure 3: Distribution of #utterances in chat context (w.r.t. the #training examples for each case).

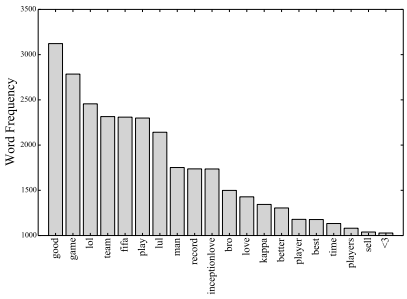

Figure 4: Frequent words in our Twitch-FIFA dataset.

4 Models

Let $v = \{v_1, v_2, .., v_m\}$ be the video context frames, $u = \{u_1, u_2, .., u_n\}$ be the textual chat (utterance) context tokens, and $r = \{r_1, r_2, .., r_k\}$ be response tokens generated (or retrieved).

4.1 Baselines

Our simple non-trained baselines are Most-Frequent-Response (re-rank the candidate responses based on their frequency in the training set), Chat-Response-Cosine (re-rank the candidate responses based on their similarity score w.r.t. the chat context), and Nearest-Neighbor (find the K-best similar chat contexts in the training set, take their corresponding responses, and then re-rank the candidate responses based on mean similarity score w.r.t. this K-best response set). For trained baselines, we use logistic regression and Naive Bayes methods. We use the final state of a Twitch-trained RNN Language Model to represent the chat context and response. Please see supplementary for full details.

4.2 Discriminative Models

4.2.1 Triple Encoder

For our simpler discriminative model, we use a 'triple encoder' to encode the video context, chat context, and response (see Fig. 5), as an extension of the dual encoder model in Lowe et al. (2015). The task here is to predict the given train-

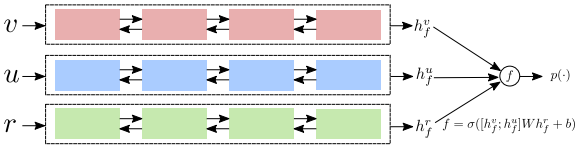

Figure 5: Overview of our 'triple encoder' discriminative model, with bidirectional-LSTM-RNN encoders for video, chat context, and response.

ing triple (v, u, r) as positive or negative. Let h_f^v, h_f^u, and h_f^r be the final state information of the video, chat, and response LSTM-RNN (bidirectional) encoders respectively; then the probability of a positive training triple is defined as follows:

$$p(v, u, r; \theta) = \sigma([h_f^v; h_f^u]^T W h_f^r + b) \quad (1)$$

where W and b are trainable parameters. Here, W can be viewed as a similarity matrix which will bring the context $[h_f^v; h_f^u]$ into the same space as the response h_f^r, and get a suitable similarity score.

For optimizing our discriminative model, we use max-margin loss function similar to Mao et al. (2016) and Yu et al. (2017). Given a positive training triple (v, u, r), let the corresponding negative training triples be (v', u, r), (v, u', r), and (v, u, r'), i.e., one modality is wrong at a time in each of these three (see Sec. 5 for the negative example selection). The max-margin loss is:

$$\begin{aligned} L(\theta) = \sum [&\max(0, M + \log p(v', u, r) - \log p(v, u, r)) \\ &+ \max(0, M + \log p(v, u', r) - \log p(v, u, r)) \\ &+ \max(0, M + \log p(v, u, r') - \log p(v, u, r))] \end{aligned} \quad (2)$$

where the summation is over all the training triples in the dataset. M is a tunable margin hyperparameter between positive and negative training triples.

4.2.2 Tridirectional Attention Flow (TriDAF)

Our tridirectional attention flow model learns stronger joint spaces between the three modalities in a mutual-information way. We use bidirectional attention flow mechanisms (Seo et al., 2017) between the video and chat contexts, between the video context and the response, as well as between the chat context and the response, hence enabling attention flow across all three modalities, as shown in Fig. 6. We name this model Tridirectional Attention Flow or TriDAF. We will next discuss the bidirectional attention flow mechanism between video and chat contexts, but the same formulation holds true for bidirectional attention between video context and response, and between chat context and response. Given the video context hidden

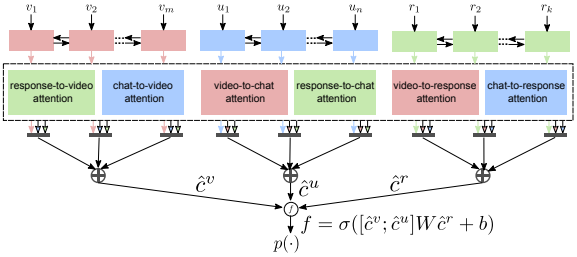

Figure 6: Overview of our tridirectional attention flow (TriDAF) model with all pairwise modality attention modules, as well as self-attention on video context, chat context, and response as inputs.

state h_i^v and chat context hidden state h_j^u at time steps i and j respectively, the bidirectional attention mechanism is based on the similarity score:

$$S_{i,j}^{(v,u)} = w_{S^{(v,u)}}^T [h_i^v; h_j^u; h_i^v \odot h_j^u] \quad (3)$$

where $S_{i,j}^{(v,u)}$ is a scalar, $w_{S^{(v,u)}}$ is a trainable parameter, and $\odot$ denote element-wise multiplication. The attention distribution from chat context to video context is defined as $\alpha_{i:} = softmax(S_{i:})$, hence the chat-to-video context vector $c_i^{v \leftarrow u} = \sum_j \alpha_{i,j} h_j^u$. Similarly, the attention distribution from video context to chat context is defined as $\beta_{j:} = softmax(S_{:j})$, hence the video-to-chat context vector $c_j^{u \leftarrow v} = \sum_i \beta_{j,i} h_i^v$.

We then compute similar bidirectional attention flow mechanisms between the video context and response, and between the chat context and response. Then, we concatenate each hidden state and its corresponding context vector from other two modalities, e.g., $\hat{h}_i^v = [h_i^v; c_i^{v \leftarrow u}; c_i^{v \leftarrow r}]$ for the i^{th} timestep of the video context. Finally, we add self-attention mechanism (Lin et al., 2017) across the concatenated hidden states of each of the three modules.[6] If $\hat{h}_i^v$ is the final concatenated vector of the video context at time step i, then the self-attention weights α^s for this video context are the softmax of e^s:

$$e_i^s = V_a^v \tanh(W_a^v \hat{h}_i^v + b_a^v) \quad (4)$$

where V_a^v, W_a^v, and b_a^v are trainable self-attention parameters. The final representation vector of the full video context after self-attention is $\hat{c}^v = \sum_i \alpha_i^s \hat{h}_i^v$. Similarly, the final representation vectors of the chat context and the response are $\hat{c}^u$ and $\hat{c}^r$, respectively. Finally, the probability that

[6] In our preliminary experiments, we found that adding self-attention is 0.92% better in recall@1 and faster than passing the hidden states through another layer of RNN, as done in Seo et al. (2017).

the given training triple (v, u, r) is positive is:

$$p(v, u, r; \theta) = \sigma([\hat{c}^v; \hat{c}^u]^T W \hat{c}^r + b) \quad (5)$$

Again, here also we use max-margin loss (Eqn. 2).

4.3 Generative Models

4.3.1 Seq2seq with Attention

Our simpler generative model is a sequence-to-sequence model with bilinear attention mechanism (similar to Luong et al. (2015)). We have two encoders, one for encoding the video context and another for encoding the chat context, as shown in Fig. 7. We combine the final state information from both encoders and give it as initial state to the response generation decoder. The two encoders and the decoder are all two-layer LSTM-RNNs. Let h_i^v and h_j^u be the hidden states of video and chat encoders at time step i and j respectively. At each time step t of the decoder with hidden state h_t^r, the decoder attends to parts of video and chat encoders and uses the combined information to generate the next token. Let α_t and β_t be the attention weight distributions for video and chat encoders respectively with video context vector $c_t^v = \sum_i \alpha_{t,i} h_i^v$ and chat context vector $c_t^u = \sum_j \beta_{t,j} h_j^u$. The attention distribution for video encoder is defined as (and the same holds for chat encoder):

$$e_{t,i} = h_t^{rT} W_a^v h_i^v; \quad \alpha_t = \text{softmax}(e_t) \quad (6)$$

where W_a^v is a trainable parameter. Next, we concatenate the attention-based context information (c_t^v and c_t^u) and decoder hidden state (h_t^r), and do a non-linear transformation to get the final hidden state $\hat{h}_t^r$ as follows:

$$\hat{h}_t^r = \tanh(W_c[c_t^v; c_t^u; h_t^r]) \quad (7)$$

where W_c is again a trainable parameter. Finally, we project the final hidden state information to vocabulary size and give it as input to a *softmax* layer to get the vocabulary distribution $p(r_t | r_{1:t-1}, v, u; \theta)$. During training, we minimize the cross-entropy loss defined as follows:

$$L_{\text{XE}}(\theta) = -\sum \sum_t \log p(r_t | r_{1:t-1}, v, u; \theta) \quad (8)$$

where the final summation is over all the training triples in the dataset.

Further, to train a stronger generative model with negative training examples (which teaches the model to give higher generative decoder probability to the positive response as compared to all the negative ones), we use a max-margin loss (similar to Eqn. 2 in Sec. 4.2.1):

$$\begin{aligned} L_{\text{MM}}(\theta) = \sum [&\max(0, M + \log p(r|v', u) - \log p(r|v, u)) \\ &+ \max(0, M + \log p(r|v, u') - \log p(r|v, u)) \\ &+ \max(0, M + \log p(r'|v, u) - \log p(r|v, u))] \end{aligned} \quad (9)$$

where the summation is over all the training triples in the dataset. Overall, the final joint loss function is a weighted combination of cross-entropy loss and max-margin loss: $L(\theta) = L_{\text{XE}}(\theta) + \lambda L_{\text{MM}}(\theta)$, where λ is a tunable hyperparameter.

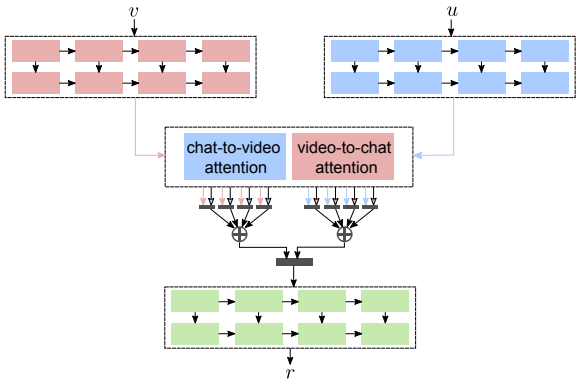

Figure 7: Overview of our generative model with bidirectional attention flow between video context and chat context during response generation.

4.3.2 Bidirectional Attention Flow (BiDAF)

The stronger version of our generative model extends the two-encoder-attention-decoder model above to add bidirectional attention flow (BiDAF) mechanism (Seo et al., 2017) between video and chat encoders, as shown in Fig. 7. Given the hidden states h_i^v and h_j^u of video and chat encoders at time step i and j, the final hidden states after the BiDAF are $\hat{h}_i^v = [h_i^v; c_i^{v \leftarrow u}]$ and $\hat{h}_j^u = [h_i^u; c_j^{u \leftarrow v}]$ (similar to as described in Sec. 4.2.2), respectively. Now, the decoder attends over these final hidden states, and the rest of the decoder process is similar to Sec 4.3.1 above, including the weighted joint cross-entropy and max-margin loss.

5 Experimental Setup

Evaluation We first evaluate both our discriminative and generative models using retrieval-based recall@k scores, which is a concrete metric for such dialogue generation tasks (Lowe et al., 2015). For our discriminative models, we simply rerank the given responses (in a candidate list of size 10, based on 9 negative examples; more details below)

Models	r@1	r@2	r@5
BASELINES			
Most-Frequent-Response	10.0	16.0	20.9
Naive Bayes	9.6	20.9	51.5
Logistic Regression	10.8	21.8	52.5
Nearest Neighbor	11.4	22.6	53.2
Chat-Response-Cosine	11.4	22.0	53.2
DISCRIMINATIVE MODEL			
Dual Encoder (C)	17.1	30.3	61.9
Dual Encoder (V)	16.3	30.5	61.1
Triple Encoder (C+V)	18.1	33.6	68.5
TriDAF+Self Attn (C+V)	20.7	35.3	69.4
GENERATIVE MODEL			
Seq2seq +Attn (C)	14.8	27.3	56.6
Seq2seq +Attn (V)	14.8	27.2	56.7
Seq2seq + Attn (C+V)	15.7	28.0	57.0
Seq2seq + Attn + BiDAF (C+V)	16.5	28.5	57.7

Table 3: Performance of our baselines, discriminative models, and generative models for recall@k metrics on our Twitch-FIFA test set. C and V represent chat and video context, respectively.

in the order of the probability score each response gets from the model. If the positive response is within the top-k list, then the recall@k score is 1, otherwise 0, following previous Ubuntu-dialogue work (Lowe et al., 2015). For the generative models, we follow a similar approach, but the reranking score for a candidate response is based on the log probability score given by the generative models' decoder for that response, following the setup of previous visual-dialog work (Das et al., 2017). In our experiments, we use recall@1, recall@2, and recall@5 scores. For completeness, we also report the phrase-matching metric scores: METEOR (Denkowski and Lavie, 2014) and ROUGE (Lin, 2004) for our generative models. We also present human evaluation.

Training Details For negative samples, during training, for every positive triple (video, chat, response) in the training set, we sample 3 random negative triples. For validation/test, we sample 9 random negative responses elsewhere from the validation/test set. Also, the negative samples don't come from the video corresponding to the positive response. More details of negative samples and other training details (e.g., dimension/vocab sizes, visual feature details, validation-based hyperparamater tuning and model selection), are discussed in the supplementary.

6 Results and Analysis

6.1 Human Evaluation of Dataset

First, the overall human quality evaluation of our dataset (shown in Table 1) demonstrates that it contains 90% responses relevant to video and/or chat context. Next, we also do a blind human study on the recall-based setup (on a set of 100 samples from the validation set), where we anonymize the positive response by randomly mixing it with 9 tricky negative responses in the retrieval list, and ask the user to select the most relevant response for the given video and/or chat context. We found that human performance on this task is around 55% recall@1, demonstrating that this 10-way-discriminative recall-based task setup is reasonably challenging for humans,[7] but also that there is a lot of scope for future model improvements because the chance baseline is only 10% and the best-performing model so far (see Sec. 6.3) achieves only 22% recall@1 (on dev set), and hence there is a large 33% gap.

6.2 Baseline Results

Table 3 displays all our primary results. We first discuss results of our simple non-trained and trained baselines (see Sec. 4.1). The 'Most-Frequent-Response' baseline, which just ranks the 10-sized response retrieval list based on their frequency in the training data, gets only around 10% recall@1.[8] Our other non-trained baselines: 'Chat-Response-Cosine' and 'Nearest Neighbor', which ranks the candidate responses based on (Twitch-trained RNN encoder's vector) cosine similarity with chat-context and K-best training contexts' response vectors, respectively, achieves slightly better scores. We also show that our simple trained baselines (logistic regression and nearest neighbor) also achieve relatively low scores, indicating that a simple, shallow model will not work on this challenging dataset.

6.3 Discriminative Model Results

Next, we present the recall@k retrieval performance of our various discriminative models in Ta-

[7] This relatively low human recall@1 performance is because this is a challenging, 10-way-discriminative evaluation, i.e., the choice comes w.r.t. 9 tricky negative examples along with just 1 positive example (hence chance-baseline is only 10%). Note that these negative examples are an artifact of specifically recall-based evaluation only, and will not affect the more important real-world task of response generation (for which our dataset's response quality is 90%, as shown in Table 1). Moreover, our dataset filtering (see Sec. 3.1) also 'suppresses' simple baselines and makes the task even harder.

[8] Note that the performance of this baseline is worse than the random choice baseline (recall@1:10%, recall@2:20%, recall@5:50%) because our dataset filtering process already suppresses frequent responses (see Sec. 3.1), in order to provide a challenging dataset for the community.

Models	METEOR	ROUGE-L
MULTIPLE REFERENCES		
Seq2seq + Atten. (C)	2.59	8.44
Seq2seq + Atten. (V)	2.66	8.34
Seq2seq + Atten. (C+V) $\otimes$	3.03	8.84
$\otimes$ + BiDAF (C+V)	3.70	9.82

Table 4: Performance of our generative models on phrase matching metrics.

Models	Relevance
Seq2seq + Atten. (C+V) wins	41.0 %
BiDAF wins	34.0 %
Non-distinguishable	25.0 %

Table 5: Human evaluation comparing the baseline and BiDAF generative models.

ble 3: dual encoder (chat context only), dual encoder (video context only), triple encoder, and TriDAF model with self-attention. Our dual encoder models are significantly better than random choice and all our simple baselines above, and further show that they have complementary information because using both of them together (in 'Triple Encoder') improves the overall performance of the model. Finally, we show that our novel TriDAF model with self-attention performs significantly better than the triple encoder model.[9]

6.4 Generative Model Results

Next, we evaluate the performance of our generative models with both retrieval-based recall@k scores and phrase matching-based metrics as discussed in Sec. 5 (as well as human evaluation). We first discuss the retrieval-based recall@k results in Table 3. Starting with a simple sequence-to-sequence attention model with video only, chat only, and both video and chat encoders, the recall@k scores are better than all the simple baselines. Moreover, using both video+chat context is again better than using only one context modality. Finally, we show that the addition of the bidirectional attention flow mechanism improves the performance in all recall@k scores.[10] Note that generative model scores are lower than the discriminative models on retrieval recall@k metric, which is expected (see discussion in previous visual dialogue work (Das et al., 2017)), because discriminative models can tune to the biases in the response candidate options, but generative models are more useful for real-world tasks such as

[9]Statistical significance of $p < 0.01$ for recall@1, based on the bootstrap test (Noreen, 1989; Efron and Tibshirani, 1994) with 100K samples.

[10]Stat. signif. $p < 0.05$ for recall@1 w.r.t. Seq2seq+Atten (video+chat); $p < 0.01$ w.r.t. chat- and video-only models.

Models	recall@1	recall@2	recall@5
1 neg.	18.21	32.19	64.05
3 neg.	22.20	35.90	68.09

Table 6: Ablation (dev) of one vs. three negative examples for TriDAF self-attention discriminative model.

generation of novel responses word-by-word from scratch in Siri/Alexa/Cortana style applications (whereas discriminative models can only rank the pre-given list of responses).

We also evaluate our generative models with phrase-level matching metrics: METEOR and ROUGE-L, as shown in Table 4. Again, our BiDAF model is stat. significantly better than non-BiDAF model on both METEOR ($p < 0.01$) and ROUGE-L ($p < 0.02$) metrics. Since dialogue systems can have several diverse, non-overlapping valid responses, we consider a multi-reference setup where all the utterances in the 10-sec response window are treated as valid responses.[11]

6.5 Human Evaluation of Models

Finally, we also perform human evaluation to compare our top two generative models, i.e., the video+chat seq2seq with attention and its extension with BiDAF (Sec. 4.3), based on a 100-sized sample. We take the generated response from both these models, and randomly shuffle these pairs to anonymize model identity. We then ask two annotators (for 50 task instances each) to score the responses of these two models based on relevance. Note that the human evaluators were familiar with Twitch FIFA-18 video games and also the Twitch's unique set of chat mannerisms and emotes. As shown in Table 5, our BiDAF based generative model performs better than the non-BiDAF one, which is already quite a strong video+chat encoder model with attention.

7 Ablations and Analysis

7.1 Negative Training Pairs

We also compare the effect of different negative training triples that we discussed in Sec. 5. Table 6 shows the comparison between one negative

[11]Liu et al. (2016b) discussed that BLEU and most phrase matching metrics are not good for evaluating dialogue systems. Also, generative models have very low phrase-matching metric scores because the generated response can be valid but still very different from the ground truth reference (Lowe et al., 2015; Liu et al., 2016b; Li et al., 2016). We present results for the relatively better metrics like paraphrase-enabled METEOR for completeness, but still focus on retrieval recall@k and human evaluation.

Figure 8: Output retrieval (left) and generative (right) examples from TriDAF and BiDAF models, resp.

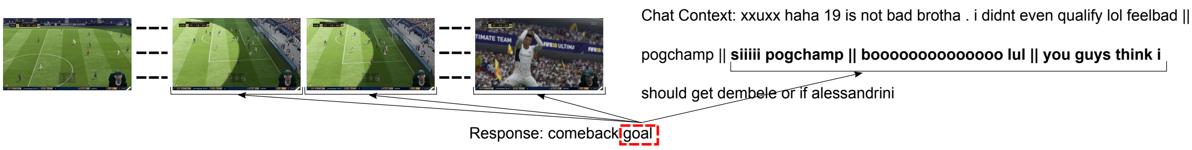

Figure 9: Attention visualization: generated word 'goal' in response is intuitively aligning to goal-related video frames (top-3-weight frames highlighted) and context words (top-10-weight words highlighted).

training triple (with just a negative response) vs. three negative training triples (one with negative video context, one with negative chat context, and another with negative response), showing that using the 3-negative examples setup is substantially better.

7.2 Discriminative Loss Functions

Table 7 shows the performance comparison between the classification loss and max-margin loss on our TriDAF with self-attention discriminative model (Sec. 4.2.2). We observe that max-margin loss performs better than the classification loss, which is intuitive because max-margin loss tries to differentiate between positive and negative training example triples.

Models	recall@1	recall@2	recall@5
Classification loss	19.32	33.72	66.60
Max-margin loss	22.20	35.90	68.09

Table 7: Ablation of classification vs. max-margin loss on our TriDAF discriminative model (on dev set).

7.3 Generative Loss Functions

For our best generative model (BiDAF), Table 8 shows that using a joint loss of cross-entropy and max-margin is better than just using only cross-entropy loss optimization (Sec. 4.3.1). Max-margin loss provides knowledge about the negative samples for the generative model, hence improves the retrieval-based recall@k scores.

7.4 Attention Visualization and Examples

Finally, we show some interesting output examples from both our discriminative and generative models as shown in Fig. 8. Additionally, Fig. 9 visualizes that our models can learn some correct attention alignments from the generated output response word to the appropriate (goal-related) video frames as well as chat context words.

Models	recall@1	recall@2	recall@5
Cross-entropy (XE)	13.12	23.45	54.78
XE+Max-margin	15.61	27.39	57.02

Table 8: Ablation of cross-entropy loss vs. cross-entropy+maxmargin loss for our BiDAF-based generative model (on dev set).

8 Conclusion

We presented a new game-chat based video-context, many-speaker dialogue task and dataset. We also presented several baselines and state-of-the-art discriminative and generative models on this task. We hope that this testbed will be a good starting point to encourage future work on the challenging video-context dialogue paradigm. In future work, we plan to investigate the effects of multiple users, i.e., the multi-party aspect of this dataset. We also plan to explore advanced video features such as activity recognition, person identification, etc.

Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by DARPA YFA17-D17AP00022, ARO-YIP Award W911NF-18-1-0336, Google Faculty Research Award, Bloomberg Data Science Research Grant, and NVidia GPU awards. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2017. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Francesco Barbieri, Luis Espinosa-Anke, Miguel Ballesteros, Horacio Saggion, et al. 2017. Towards the understanding of gaming audiences by modeling twitch emotes. In *Third Workshop on Noisy User-generated Text (W-NUT 2017)*.

Madeleine Bates. 1995. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982.

Justine Cassell. 1999. Embodied conversation: integrating face and gesture into automatic spoken dialogue systems.

Asli Celikyilmaz, Li Deng, and Dilek Hakkani-Tur. 2017. Deep learning for spoken and text dialog systems. *Deep Learning in Natural Language Processing (eds. Li Deng and Yang Liu)*.

Asli Celikyilmaz, Zhaleh Feizollahi, Dilek Hakkani-Tur, and Ruhi Sarikaya. 2014. Resolving referring expressions in conversational dialogs for natural user interfaces. In *EMNLP*, pages 2094–2104.

Asli Celikyilmaz, Zhaleh Feizollahi, Dilek Hakkani-Tur, and Ruhi Sarikaya. 2015. A universal model for flexible item selection in conversational dialogs. In *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 361–367. IEEE.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM.

Praveen Damacharla, Parashar Dhakal, Sebastian Stumbo, Ahmad Y Javaid, Subhashini Ganapathy, David A Malek, Douglas C Hodge, and Vijay Devabhaktuni. 2018. Effects of voice-based synthetic assistant on performance of emergency care provider in training. *International Journal of Artificial Intelligence in Education*, pages 1–22.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *CVPR*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Renato De Mori. 2007. Spoken language understanding: a survey. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 365–376. IEEE.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

John Dowding, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. 1993. Gemini: A natural language system for spoken-language understanding. In *ACL*, pages 54–61.

Wieland Eckert, Esther Levin, and Roberto Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 80–87. IEEE.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Mary Ellen Foster, Ellen Gurman Bard, Markus Guhe, Robin L Hill, Jon Oberlander, and Alois Knoll. 2008. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pages 295–302. ACM.

Cheng-Yang Fu, Joon Lee, Mohit Bansal, and Alexander C Berg. 2017. Video highlight prediction using audience chat reactions. In *EMNLP*.

Kallirroi Georgila, James Henderson, and Oliver Lemon. 2006. User simulation for spoken dialogue systems: Learning and evaluation. In *Ninth International Conference on Spoken Language Processing*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2017. Iqa: Visual question answering in interactive environments. *arXiv preprint arXiv:1712.03316*.

Iryna Gurevych and Michael Strube. 2004. Semantic similarity applied to spoken dialogue summarization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 764. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Takaaki Hori, Hai Wang, Chiori Hori, Shinji Watanabe, Bret Harsham, Jonathan Le Roux, John R Hershey, Yusuke Koji, Yi Jing, Zhaocheng Zhu, et al. 2016. Dialog state tracking with attention-based sequence-to-sequence learning. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 552–558. IEEE.

Charles Lee Isbell, Michael Kearns, Dave Kormann, Satinder Singh, and Peter Stone. 2000. Cobot in lambdamoo: A social statistics agent. In *AAAI/IAAI*, pages 36–41.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2680–8.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–I. IEEE.

Michael Johnston, Srinivas Bangalore, Gunaranjan Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, and Preetam Maloor. 2002. Match: An architecture for multimodal dialogue systems. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 376–383. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *ICLR*.

Changsong Liu, Joyce Y Chai, Nishant Shukla, and Song-Chun Zhu. 2016a. Task learning through visual demonstration and situated dialogue. In *AAAI Workshop: Symbiotic Cognitive Systems*.

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

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Yi Luan, Yangfeng Ji, and Mari Ostendorf. 2016. Lstm based conversation models. *arXiv preprint arXiv:1603.09457*.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *CVPR*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Nikola Mrkšić, Diarmuid O Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2016. Neural belief tracker: Data-driven dialogue state tracking. *arXiv preprint arXiv:1606.03777*.

Karen Myers, Pauline Berry, Jim Blythe, Ken Conley, Melinda Gervasio, Deborah L McGuinness, David Morley, Avi Pfeffer, Martha Pollack, and Milind Tambe. 2007. An intelligent personal assistant for task and time management. *AI Magazine*, 28(2):47.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Qing Ping and Chaomei Chen. 2017. Video highlights detection and summarization with lag-calibration based on concept-emotion mapping of crowd-sourced time-sync comments. In *EMNLP Workshop on New Frontiers in Summarization*.

Owen Rambow, Srinivas Bangalore, and Marilyn Walker. 2001. Natural language generation in dialog systems. In *Proceedings of the first international conference on Human language technology research*, pages 1–4. Association for Computational Linguistics.

Verena Rieser, Ivana Kruijff-Korbayová, and Oliver Lemon. 2005. A corpus collection and annotation framework for learning multimodal clarification strategies. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*.

Verena Rieser and Oliver Lemon. 2008. Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. In *ACL*, pages 638–646.

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

Ruhi Sarikaya, Paul A Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Celikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, Xiaohu Liu, et al. 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pages 391–397. IEEE.

Kjeld Schmidt and Liam Bannon. 1992. Taking cscw seriously. *Computer Supported Cooperative Work (CSCW)*, 1(1-2):7–40.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron C Courville. 2017a. Multiresolution recurrent neural networks: An application to dialogue response generation. In *AAAI*, pages 3288–3294.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Pararth Shah, Dilek Hakkani-Tür, and Larry Heck. 2016. Interactive reinforcement learning for task-oriented dialogue management. In *NIPS Deep Learning for Action and Interaction Workshop*.

Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 2000. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*, pages 956–962.

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. *arXiv preprint arXiv:1506.06714*.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *ACL*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *CVPR*, pages 4534–4542.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of ICML Deep Learning Workshop*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner. 2001. Advances in automatic meeting record creation and access. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, pages 597–600. IEEE.

Joseph Weizenbaum. 1966. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

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*.

Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. 2018. CHALET: Cornell house agent learning environment. *arXiv preprint arXiv:1801.07357*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Steve J Young. 2000. Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1389–1402.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*.