Overview of the seventh Dialog System Technology Challenge: DSTC7

D'Haro, Luis Fernando; Yoshino, Koichiro; Hori, Chiori; Marks, Tim; Polymenakos, Lazaros; Kummerfeld, Jonathan K.; Galley, Michel; Gao, Xiang

TR2020-029 March 18, 2020

Abstract

This paper provides detailed information about the seventh Dialog System Technology Challenge (DSTC7) and its three tracks aimed to explore the problem of building robust and accurate end-to-end dialog systems. In more detail, DSTC7 focuses on developing and exploring end-to-end technologies for the following three pragmatic challenges: (1) sentence selection for multiple domains, (2) generation of informational responses grounded in external knowledge, and (3) audio visual scene-aware dialog to allow conversations with users about objects and events around them. This paper summarizes the overall setup and results of DSTC7, including detailed descriptions of the different tracks, provided datasets and annotations, overview of the submitted systems and their final results. For Track 1, LSTM-based models performed best across both datasets, allowing teams to effectively handle task variants where no correct answer was present or when multiple paraphrases were included. For Track 2, RNN-based architectures augmented to incorporate facts by using two types of encoders: a dialog encoder and a fact encoder plus using attention mechanisms and a pointer-generator approach provided the best results. Finally, for Track 3, the best model used Hierarchical Attention mechanisms to combine the text and vision information obtaining a 22% better result than the baseline LSTM system for the human rating score. More than 220 participants were registered and about 40 teams participated in the final challenge. 32 scientific papers reporting the systems submitted to DSTC7, and 3 general technical papers for dialog technologies, were presented during the one-day wrap-up workshop at AAAI-19. During the workshop, we reviewed the state-of-the-art systems, shared novel approaches to the DSTC7 tasks, and discussed the future directions for the challenge (DSTC8).

Computer Speech and Language

Overview of the seventh Dialog System Technology Challenge: DSTC7

Luis Fernando D'Haro[a], Koichiro Yoshino[b], Chiori Hori[c], Tim K. Marks[c], Lazaros Polymenakos[d], Jonathan K. Kummerfeld[e], Michel Galley[f], Xiang Gao[f,1]

[a]*Speech Technology Group. Center for Information Processing and Telecommunications (IPTC), ETSI Telecomunicación Universidad Politécnica de Madrid, Ciudad Universitaria, Av. Complutense, 30, 28040 Madrid, Spain*
[b]*Nara Institute of Science and Technology, Ikoma, Nara, 6300192, Japan*
[c]*Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA, 02139, USA*
[d]*Alexa Dialog Science, 101 Main Street, Cambridge, MA, 02142, USA*
[e]*University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, USA*
[f]*Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA*

Abstract

This paper provides detailed information about the seventh Dialog System Technology Challenge (DSTC7) and its three tracks aimed to explore the problem of building robust and accurate end-to-end dialog systems. In more detail, DSTC7 focuses on developing and exploring end-to-end technologies for the following three pragmatic challenges: (1) sentence selection for multiple domains, (2) generation of informational responses grounded in external knowledge, and (3) audio visual scene-aware dialog to allow conversations with users about objects and events around them.

This paper summarizes the overall setup and results of DSTC7, including detailed descriptions of the different tracks, provided datasets and annotations, overview of the submitted systems and their final results. For Track 1, LSTM-based models performed best across both datasets, allowing teams to effectively handle task variants where no correct answer was present or when multiple paraphrases were included. For Track 2, RNN-based architectures augmented to incorporate facts by using two types of encoders: a dialog encoder and a fact encoder plus using attention mechanisms and a

[1]Every author has equal contribution. http://workshop.colips.org/dstc7

pointer-generator approach provided the best results. Finally, for Track 3, the best model used Hierarchical Attention mechanisms to combine the text and vision information obtaining a 22% better result than the baseline LSTM system for the human rating score.

More than 220 participants were registered and about 40 teams participated in the final challenge. 32 scientific papers reporting the systems submitted to DSTC7, and 3 general technical papers for dialog technologies, were presented during the one-day wrap-up workshop at AAAI-19. During the workshop, we reviewed the state-of-the-art systems, shared novel approaches to the DSTC7 tasks, and discussed the future directions for the challenge (DSTC8).

DSTC7: Dialog Challenge to build more robust and accurate end-to-end dialog systems.

Track 1, Sentence selection for multiple domains, including variations where there are a large number of candidate options, and where the candidate set has zero, one, or multiple correct options.

Track 2, Beyond Chitchat: Generation of informational responses grounded in external knowledge.

Track 3, Audio visual scene-aware dialog systems to allow dynamic conversations about objects and events around users.

Manuscript File

Overview of the seventh Dialog System Technology Challenge: DSTC7

Luis Fernando D'Haro[a], Koichiro Yoshino[b], Chiori Hori[c], Tim K. Marks[c], Lazaros Polymenakos[d], Jonathan K. Kummerfeld[e], Michel Galley[f], Xiang Gao[f,1]

[a]*Speech Technology Group. Center for Information Processing and Telecommunications (IPTC), ETSI Telecomunicación Universidad Politécnica de Madrid, Ciudad Universitaria, Av. Complutense, 30, 28040 Madrid, Spain*
[b]*Nara Institute of Science and Technology, Ikoma, Nara, 6300192, Japan*
[c]*Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA, 02139, USA*
[d]*Alexa Dialog Science, 101 Main Street, Cambridge, MA, 02142, USA*
[e]*University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, USA*
[f]*Microsoft Research, One Microsoft Way, Redmond, WA, 98052, USA*

Abstract

This paper provides detailed information about the seventh Dialog System Technology Challenge (DSTC7) and its three tracks aimed to explore the problem of building robust and accurate end-to-end dialog systems. In more detail, DSTC7 focuses on developing and exploring end-to-end technologies for the following three pragmatic challenges: (1) sentence selection for multiple domains, (2) generation of informational responses grounded in external knowledge, and (3) audio visual scene-aware dialog to allow conversations with users about objects and events around them.

This paper summarizes the overall setup and results of DSTC7, including detailed descriptions of the different tracks, provided datasets and annotations, overview of the submitted systems and their final results. For Track 1, LSTM-based models performed best across both datasets, allowing teams to effectively handle task variants where no correct answer was present or when multiple paraphrases were included. For Track 2, RNN-based architectures augmented to incorporate facts by using two types of encoders: a dialog encoder and a fact encoder plus using attention mechanisms and a pointer-generator approach provided the best results. Finally, for Track 3, the best model used Hierarchical Attention mechanisms to combine the text

[1]Every author has equal contribution. http://workshop.colips.org/dstc7

and vision information obtaining a 22% better result than the baseline LSTM system for the human rating score.

More than 220 participants were registered and about 40 teams participated in the final challenge. 32 scientific papers reporting the systems submitted to DSTC7, and 3 general technical papers for dialog technologies, were presented during the one-day wrap-up workshop at AAAI-19. During the workshop, we reviewed the state-of-the-art systems, shared novel approaches to the DSTC7 tasks, and discussed the future directions for the challenge (DSTC8).

Keywords:
Dialog System Technology Challenge, end-to-end dialog systems, Sentence Selection, Natural Language Generation, Audio Visual Scene-Aware Dialog.

1. Introduction

The ongoing DSTC series started as an initiative to provide a common testbed for the task of Dialog State Tracking; the first edition was organized in 2013 (Williams et al. (2013)) and used human-computer dialogs in the bus timetable domain. Dialog State Tracking Challenges 2 (Henderson et al. (2014a)) and 3 (Henderson et al. (2014b)) followed in 2014, using more complicated and dynamic dialog states for restaurant information in different situations, e.g. state tracking for unseen states, and tested with different domain data. Dialog State Tracking Challenge 4 (Kim et al. (2017)) and Dialog State Tracking Challenge 5 (Kim et al. (2016)) moved to tracking human-human dialogs in mono- and cross-language settings. Then, for DSTC6 in 2017, the challenge focused on end-to-end systems with the aim of minimizing effort on human annotation while exploring more complex and diverse tasks related with dialog systems (Hori et al. (2019c)). For this last edition, DSTC7 in 2018, we focused on scaling the capabilities of the systems, explore multimodal approaches and better use of external information.

It is clear that, since its first edition in 2013, the challenge has evolved in several ways. First, from modeling human-computer interactions, then to explore human-human interactions, and finally moving toward complex and more robust end-to-end systems. DSTC has also offered pilot tasks on speech act prediction, spoken language understanding, natural language generation, and end-to-end system evaluation, which expanded interest in the challenge for the dialog and AI research communities. Therefore, given the remarkable success of the first five editions, the complexity of the dialog phenomenon and the interest of the research community in the broader variety of dialog related problems, the DSTC rebranded itself as “Dialog System Technology Challenges” since its sixth edition.

For the seventh edition, there were five task proposals. These were discussed during the AAAI-19 workshop, with a focus on how applied proposals were, and how they fit within the larger space of problems of interest to the research community. Three critical issues were raised in the discussion. First, despite the enormous success of the generative approaches used in neural conversation models for response generation, retrieval-based approaches are still essential from a practical point of view (Sentence Selection Track). Second, improving generative approaches is important too in order to allow more response variety considering the dialog context, dialog history, other dialog situations, and grounding the responses by means of external knowledge (Sentence Generation Track). The final issue was to extend the dialog systems with complementary multimodal information to allow the system to understand better the context, and allowing the fusion with other research areas; visual dialog is one direction in which information in images is used in the dialog (Audio Visual Scene-Aware Dialog Track). Following this discussion, three tasks were selected for the seventh Dialog System Technology Challenge, as described below.

For the Sentence Selection track (described in more detail in section 2), the challenge consists of five sub-tasks, in which systems are given a partial conversation, and they must select the correct next utterance from a short or very large set of candidates, including paraphrases as candidates, or indicate that none of the proposed utterances is correct. This is intended to push the utterance classification task towards real-world problems.

For the Sentence Generation track (described in detail in section 3), the goal is to generate informative responses that go beyond chitchat, in this case by injecting informational responses that are grounded in external knowledge (e.g., news stories, or background information such as Wikipedia pages). This task is indented to promote research on fully data-driven response generation—which has so far been mostly limited to chitchat—by combining the benefits of fully end-to-end approaches with more practical purposes (e.g., informing the users rather than just entertaining them).

Finally, in the Audio Visual Scene-aware Dialog track (described in detail in section 4), the goal is to generate system responses in a dialog about an input video. Dialog systems need to understand scenes to have conversations with users about the objects and events around them. In this track, multiple research technologies are integrated including: end-to-end dialog technologies, which generate system responses using models trained from dialog data; visual question answering (VQA) technologies, which answer to questions about images using learned image features; and video description technologies, in which videos are described/narrated using multimodal information.

1.1. Workshop summary and future DSTC

The workshop for the Dialog System Technology Challenge (DSTC) was held on January 27, 2019 at Honolulu, Hawaii, USA, collocated with the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). More than 220 participants were registered in one or several of the proposed three tasks; finally, about 40 teams submitted their final results and 32 scientific papers were presented during the workshop, together with 3 general technical papers about dialog systems. We had about 80 pre-registrations for the workshop and more participants joined on-site. The workshop also had many supporting organizations including three sponsors, and an invited talk about Massively Multilingual Dialog and Q&A by Dr. Holger Schwenk.

In addition, as part of our efforts to promote the research in dialog technologies, we presented the challenge, tracks, provided data and results during the 2nd NeurIPS workshop on Conversational AI: Today's Practice and Tomorrow's Potential[2].

Finally, to initiate DSTC8, from November 22, 2018 until January 11, 2019 we received up to 7 track proposals for DSTC8 [3]. During the AAAI-19 workshop these proposals were presented to the attendees and then we passed them a survey to know their interest and willingness to participate on each; after the workshop, the following tracks were selected: a) End-to-end Task Completion b) Predicting Responses, c) Audio Visual Scene-Aware Dialog, and d)Schema-Guided State Tracking. This way, we will continue focusing on end-to-end dialog tasks and their application to Dialog Systems in a pragmatic way.

[2] http://alborz-geramifard.com/workshops/nips18-Conversational-AI/Main.html

[3] For detailed information about each proposal and the selection criteria check: http://workshop.colips.org/dstc7/dstc8_proposals.html

2. Sentence Selection Track

Automatic dialogue systems have great potential as a new form of user interface between people and computers. Unfortunately, there are relatively few large resources of human-human dialogues (Serban et al., 2018), which are crucial for the development of robust statistical models. Evaluation also poses a challenge, as the output of an end-to-end dialogue system could be entirely reasonable, but not match the reference, either because it is a paraphrase, or it takes the conversation in a different, but still coherent, direction.

In this track, we introduced two new datasets and explored variations in task structure for research on goal-oriented dialogue. One of our datasets was carefully constructed with real people acting in a university student advising scenario. The other dataset was formed by applying a new disentanglement method (Kummerfeld et al., 2018) to extract conversations from an IRC channel of technical help for the Ubuntu operating system. We structured the dialogue problem as next utterance selection, in which participants receive partial dialogues and must select the next utterance from a set of options. Going beyond prior work, we considered larger sets of options, and variations with either additional incorrect options, paraphrases of the correct option, or no correct option at all. These changes push the next utterance selection task towards real-world dialogue.

This task is not a continuation of prior DSTC tasks, but it is related to tasks 1 and 2 from DSTC6 (Perez et al., 2017; Hori and Hori, 2017a). Like DSTC6 task 1, our task considers goal-oriented dialogue and next utterance selection, but our data is from human-human conversations, whereas theirs was simulated. Like DSTC6 task 2, we use online resources to build a large collection of dialogues, but their dialogues were shorter (2 - 2.5 utterances per conversation) and came from a more diverse set of sources (1,242 twitter customer service accounts, and a range of films).

Below we provide an overview of (1) the task structure, (2) the datasets, (3) the evaluation metrics, and (4) system results. Twenty teams participated, with one clear winner, scoring the highest on all but one sub-task. The data and other resources associated with the task have been released[4] to enable future work on this topic and to make accurate comparisons possible.

[4]https://ibm.github.io/dstc7-noesis/public/index.html

2.1. Task

This task pushed the state-of-the-art in goal-oriented dialogue systems in four directions deemed necessary for practical automated agents, using two new datasets. We sidestepped the challenge of evaluating generated utterances by formulating the problem as next utterance selection, as proposed by Lowe et al. (2015). At test time, participants were provided with partial conversations, each paired with a set of utterances that could be the next utterance in the conversation. Systems needed to rank these options, with the goal of placing the true utterance first. Prior work used sets of 2 or 10 utterances. We make the task harder by expanding the size of the sets, and considered several advanced variations:

Subtask 1 100 candidates, including 1 correct option.

Subtask 2 120,000 candidates, including 1 correct option (Ubuntu data only).

Subtask 3 100 candidates, including 1-5 correct options that are paraphrases (Advising data only).

Subtask 4 100 candidates, including 0-1 correct options.

Subtask 5 The same as subtask 1, but with access to external information.

These subtasks push the capabilities of systems. In particular, when the number of candidates is small (2-10) and diverse, it is possible that systems are learning to differentiate topics rather than learning dialogue. Our variations move towards a task that is more representative of the challenges involved in dialogue modeling.

As part of the challenge, we provided a baseline system that implemented the Dual-Encoder model from Lowe et al. (2015). This lowered the barrier to entry, encouraging broader participation in the task.

2.2. Data

We used two datasets containing goal-oriented dialogues between two participants, but from very different domains. This challenge introduced the two datasets, and we kept the test set answers secret until after the challenge.[5]

[5] The entire datasets are now publicly available at https://ibm.github.io/dstc-noesis/public/index.html

10:30	<elmaya>	is there a way to setup grub to not press the esc button for the menu choices?
10:31	<scaroo>	elmaya, edit /boot/grub/ menu.lst and comment the "hidemenu" line
10:32	<scaroo>	elmaya, then run grub -install
10:32	<scaroo>	grub-install
10:32	<elmaya>	thanls scaroo
10:32	<elmaya>	thanks

Figure 1: Example Ubuntu dialogue before our pre-processing.

To construct the partial conversations we randomly split each conversation. Incorrect candidate utterances are selected by randomly sampling utterances from the rest of the dataset. For subtask 3 (paraphrases), the incorrect candidates are sampled with paraphrases as well. For subtask 4 (no correct option sometimes), twenty percent of examples were randomly sampled and the correct utterance was replaced with an additional incorrect one.

Along with the datasets we provided additional sources of information that were specific to each dataset. Participants were able to use the provided knowledge sources as is, or automatically transform them to appropriate representations (e.g. knowledge graphs, continuous embeddings, etc.) that were integrated with end-to-end dialogue systems so as to increase response accuracy.

2.2.1. Ubuntu

We constructed one dataset from the Ubuntu Internet Relay Chat (IRC) support channel, in which users help each other to resolve technical problems related to the Ubuntu operating system. We consider only conversations in which one user asks a question and another helps them resolve their problem. We extracted conversations from the channel using the conversational disentanglement method described by Kummerfeld et al. (2018), trained with manually annotated data using Slate (Kummerfeld, 2019).[6,7] See Kummerfeld et al. (2018) for detailed analysis of the extraction process. At a high

[6]Previously, Lowe et al. (2015) extracted conversations from the same IRC logs, but with a heuristic method. Kummerfeld et al. (2018) showed that the heuristic was far less effective than a trained statistical model.

[7]The specific model used in DSTC 7 track 1 is from an earlier version of Kummerfeld et al. (2018), as described in the ArXiv preprint and released as the C++ version.

level, we used a feedforward neural network that considers each message in the logs and predicts which earlier message it is a response to. This forms a structure in which each connected component is a single conversation. The manual annotation of the data had a convention that when a user asks a question that starts a new conversation, which makes it clear who is asking for help and who is providing it.

We further applied several filters to increase the quality of the extracted dialogues: (1) the first message must not be directed, (2) there are exactly two participants (a questioner and a helper), not counting the channel bot, (3) no more than 80% of the messages are by a single participant, and (4) there are at least three turns. This approach produced 135,000 conversations, and each was cut off at different points to create the necessary conversations for all the subtasks. In all cases, the cutoff point was chosen to ensure there were at least three prior turns of dialogue.

Figure 1 shows an example dialogue from the dataset. For the actual challenge we identify the users as 'speaker_1' (the person asking the question) and 'speaker_2' (the person answering), and removed usernames from the messages (such as 'elmaya' in the example). We also combined consecutive messages from a single user, and always cut conversations off so that the last speaker was the person asking the question. This meant systems were learning to behave like the helpers, which fits the goal of developing a dialogue system to provide help.

For subtask 5, additional data was provided in the form of manual pages. These provide information on commands that are frequently mentioned in the Ubuntu technical support conversations.

2.2.2. Advising

Our second dataset is based on an entirely new collection of dialogues in which university students are being advised which classes to take. These were collected at the University of Michigan with IRB approval. Pairs of Michigan students play-acted the roles of a student and an advisor. We provided a persona for the student, describing the classes they had taken already, what year of their degree they were in, and several types of class preferences (workloads, class sizes, topic areas, time of day, etc.). Advisors did not know the student's preferences, but did know what classes they had taken, what classes were available, and which were suggested (based on aggregate statistics from real student records). The data was collected over a year, with some data collected as part of courses in NLP and social computing,

Student	Hi professor, I am looking for courses to take. Do you have any suggestions?
Advisor	What topic do you prefer, computer science or electrical engineering?
Student	I prefer electrical engineering.
Advisor	Based on your background, I would like to suggest you take one of the two courses: EECS 550 Information Theory and EECS 551: Matrix Methods for Signal Processing, Data Analysis and Machine Learning FA 2012
Student	Can you describe a little bit about EECS 550?
Advisor	This course contains a lot of concepts about source, channel, rate of transformation of information, etc.
Student	Sounds interesting. Do you know the class size of this course?
Advisor	This is a relatively small class and the average size of it is around 12.
Student	I would prefer class with larger class size. What is EECS 551 about?
Advisor	This course is about theory and application of matrix methods to signal processing, data analysis and machine learning
Student	What is the course size of EECS 551?
Advisor	It is around 71
Student	I would take EECS 551. Thanks professor!
Advisor	You are welcome!

Student	Hello, I need some courses. What would you recommend?
Student	Hi. I am looking for courses. Could you recommend some?
Student	Hi professor, what courses would you suggest for me to take?
Student	Professor, what courses do you advise I take?
Student	Hi professor, I look for suggestions to take courses. Do you have any?
Student	Hi professor, do you have any suggestions to search and take courses?
Advisor	Is electrical engineering your main interest, or computer science?
Advisor	Do you think you would like computer science or electrical engineering best?
Advisor	computer science or electrical engineering,which do you prefer?
Advisor	Do u prefer computer science or do u prefer electrical engineering?
Advisor	Which subject is more interesting to you, computer science or electrical engineering?
Advisor	Do you prefer computer science or electrical engineering?

Figure 2: Example Advising dialogue and paraphrases of the first two utterances.

Property	Advising	Ubuntu
Dialogues	815	135,078
Utterances / Dialogue	18.3	10.0
Tokens / Utterance	9.8	9.9
Utterances / Unique utt.	1.1	1.1
Tokens / Unique tokens	50.8	22.9

Table 1: Comparison of the diversity of the complete underlying datasets (train, dev, test, and unused). Advising is smaller, has longer conversations, and more token diversity. Tokens are based on splitting on whitespace.

and some collected with paid participants.

In the shared task, we provide all of this information - student preferences, and course information - to participants. 815 conversations were collected, and then the data was expanded by collecting 82,094 paraphrases using the crowdsourcing approach described by Jiang et al. (2017). This involved asking each worker for multiple paraphrases, with carefully designed examples that guided them towards creative edits that were still correct. Of this data, 500 conversations were used for training, 100 for development, and 100 for testing. The remaining 115 conversations were used to create a large pool of utterances. This pool was then used as a source of negative candidate sentences in the candidate sets. For the test data, 500 conversations were constructed by cutting the conversations off at 5 points and using paraphrases to make 5 distinct conversations. The training data was provided in two forms. First, the 500 training conversations with a list of paraphrases for each utterance, which participants could use in any way. Second, 100,000 partial conversations generated by randomly selecting paraphrases for every message in each conversation and selecting a random cutoff point.

Two versions of the test data were provided to participants. A mistake led to the first version of the test set drawing from both training and test dialogues, rather than using just the test dialogues. During the challenge this issue was identified and a corrected version was released to all participants. Results on both sets were included in the initial task summary, but we only include the final set here and encourage all future work to only consider the second test set.

2.2.3. Comparison

Table 1 provides statistics about the two raw datasets. The Ubuntu dataset is based on several orders of magnitude more conversations, but they are automatically extracted, which means there are errors (conversations that are missing utterances or contain utterances from other conversations). Both have similar length utterances, but these values are on the original Ubuntu dialogues, before we merge consecutive messages from the same user. The Advising dialogues contain more messages on average, but the Ubuntu dialogues cover a wider range of lengths (up to 118 messages). Interestingly, the diversity in tokens varies substantially, while utterance lengths and utterance diversity are similar.

2.3. Results

Twenty teams submitted entries for at least one subtask. Additional external resources were not permitted, with the exception of pre-trained embeddings that were publicly available prior to the release of the data.

2.3.1. Participants

Table 2 presents a summary of approaches teams used. One clear trend was the use of the Enhanced LSTM model (ESIM, Chen et al., 2017), though each team modified it differently as they worked to improve performance on the task. Other approaches covered a wide range of neural model components: Convolutional Neural Networks, Memory Networks, the Transformer, Attention, and Recurrent Neural Network variants. Two teams used ELMo word representations (Peters et al., 2018), while three constructed ensembles. Several teams also incorporated more classical approaches, such as TF-IDF based ranking, as part of their system.

We provided a range of data sources in the task, with the goal of enabling innovation in training methods. Six teams used the external data, while four teams used the raw form of the Advising data. The rules did not state whether the validation data could be used as additional training data at test time, and so we asked each team what they used. As Table 2 shows, only four teams trained their systems with the validation data.

2.3.2. Metrics

We considered a range of metrics when comparing models. Following Lowe et al. (2015), we use Recall@N, where we count how often the correct answer is within the top N specified by a system. In prior work, there were either 2 or 10 candidates (including the correct one), and N was set at 1, 2, or 5. Our sets are larger, with 100 candidates, and so we considered larger values of N: 1, 10, and 50. 10 and 50 were chosen to correspond to 1 and 5 in prior work (the expanded candidate set means they correspond to the same fraction of the space of options). We also considered a widely used metric from the ranking literature: Mean Reciprocal Rank (MRR). For subtask 3 we measured Mean Average Precision (MAP) since there are multiple correct utterances in the set. Finally, for subtask 4, participants had to return 101 values, the extra one being the value 'NONE', to indicate that no valid answer was present.

To determine a single winner for each subtask, we used the mean of Recall@10 and MRR, as presented in Table 3.

2.3.3. Discussion

Table 3 presents the overall scores for each team on each subtask, ordered by teams' average rank. Team 3 consistently scored highest, winning all but one subtask. For details of their approach, see Chen and Wang (2019). Looking at individual metrics, they had the best score 75% of the time on Ubuntu and all of the time on the final Advising test set. The subtask they were beaten on was Ubuntu-2, in which the set of candidates was drastically expanded. Team 10 did best on that task, indicating that their extra filtering step provided a key advantage. They filtered the 120,000 sentence set down to 100 options using a TF-IDF based method, then applied their standard approach to that set. For details of the method, see Ganhotra et al. (2019).

Subtasks.

1. The first subtask drew the most interest, with every team participating in it for one of the datasets. Performance varied substantially, covering a wide range for both datasets, particularly on Ubuntu.
2. As expected, subtask 2 was more difficult than task 1, with consistently lower results. However, while the number of candidates was increased from 100 to 120,000, performance reached as high as half the level of task 1, which suggests systems could handle the large set effectively.
3. Also as expected, results on subtask 3 were slightly higher than on subtask 1. Comparing MRR and MAP it is interesting to see that while the ranking of systems is the same, in some cases MAP was higher than MRR and in others it was lower.

4. For both datasets, results on subtask 4, where the correct answer was to choose no option 20% of the time, are generally similar. On average, no metric shifted by more than 0.016, and some went up while others went down. This suggests that teams were able to effectively handle the added challenge.
5. Finally, on subtask 5 we see some slight gains in performance, but mostly similar results, indicating that effectively using external resources remains a challenge.

Advising Test Sets. We compared results on the two versions of the test set (one which had overlap with the source dialogues from training, and the other with entirely distinct dialogues). Removing overlap made the task considerably harder, though more realistic. In general, system rankings were not substantially impacted, with the exception of team 17, which did better on the original dataset. This may relate to their use of a memory network over the raw advising data, which may have led the model to match test dialogues with their corresponding training dialogues.

Metrics. Finally, we compared the metrics. In 39% of cases a team's ranking is identical across all metrics, and in 34% there is a difference of only one place. The maximum difference is 5, which occurred once, between team 6's results in the final Advising results, where their Recall@1 result was 8th, their Recall@10 result was 11th and their Recall@50 result was 13th. Comparing MRR and Recall@N, the MRR rank is outside the range of ranks given by the recall measures 9% of the time (on Ubuntu and the final Advising evaluation).

2.4. Future Work

This task provides the basis for a range of interesting new directions. We randomly selected negative options, but other strategies could raise the difficulty, for example by selecting very similar candidates according to a simple model. For evaluation, it would be interesting to explore human judgements, since by expanding the candidate sets we are introducing options that are potentially reasonable.

This work has been extended in several direction by a follow-up task at DSTC 8. In particular, the setting was expanded to include conversations with more than two participants. One subtask also explores the challenge of selecting responses in the raw channel, where multiple conversations are occurring at once. These pose additional challenges and bring the setting closer to the real world. The data has also been improved, by using an improved version of the disentanglement algorithm that extracts higher quality conversations.

2.5. Conclusion

This task introduced two new datasets and three new variants of the next utterance selection task. Twenty teams attempted the challenge, with one clear winner. The datasets are being publicly released, along with a baseline approach, in order to facilitate further work on this task. This resource will support the development of novel dialogue systems, pushing research towards more realistic and challenging settings.

3. Sentence Generation Track

Recent work (Ritter et al., 2011; Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015; Serban et al., 2016, etc.) has shown that conversational models can be trained in a completely end-to-end and data-driven fashion, without any hand-coding. However, prior work has mostly focused to chitchat, as that is a common feature of messages in the social media data (e.g., Twitter (Ritter et al., 2011)) used to train these systems. Such end-to-end neural conversation systems have a tendency to produce responses that are conversationally appropriate, but that are also often bland (Li et al., 2016a; Gao et al., 2019), purely chatty, and lacking entities and factual content. On the other end, goal-oriented dialog systems have the ability to inject entities and facts into responses, but often at the cost of significant hand-coding (e.g., slot filling) and this hand-crafting is often specific to the domain or task. We argue that dialog shouldn't necessarily be either completely goal-oriented or completely chitchat. This is often reflected in real human-human data, which often combines the two genres.

To effectively move beyond chitchat and produce system responses that are both substantive and "useful", fully data-driven models need grounding in the real world and access to external knowledge (textual or structured). To do so, the Sentence Generation task was inspired by the *knowledge-grounded* conversational framework of (Ghazvininejad et al., 2018; Qin et al., 2019), which combines conversational input and textual data from the user's environment (here, a web page that is discussed). Such a framework maintains the benefit of fully data-driven conversation while attempting to get closer to task-oriented scenarios, with the goal of informing and helping the users and not just entertaining them.

3.1. Task definition

The task follows the data-driven framework established in 2011 by Ritter et al. (2011), which avoids hand-coding any linguistic, domain, or task-specific information (e.g., there are no explicit dialog act or slots). In the knowledge-grounded setting of (Ghazvininejad et al., 2018; Qin et al., 2019), that framework is extended as each system input consists of two parts:

- **Conversational input**: Similar to DSTC6 Track 2 (Hori and Hori, 2017b), all preceding turns of the conversation are available to the system. For practical purposes, we truncate the context to the K most recent turns.
- **Contextually-relevant "facts":** The system is given text that is relevant to the context of the conversation, in this case a web page. This text is distinct from conversational data, and is extracted from external knowledge sources such as Wikipedia or news web sites.

From this input, the task it to produce a response that is (1) conversationally appropriate and relevant, as well as (2) informative and interesting. The evaluation setup is presented in Section 3.4, which includes a human evaluation of these two qualities ("Relevance" and "Interest", respectively).

3.2. Data

We extracted conversation threads from Reddit data, which is particularly well suited for grounded conversation modeling. Indeed, Reddit conversations are organized around submissions, where each conversation is typically initiated with a URL to a web page (grounding) that defines the subject of the conversation. An example of the data is shown in Table 4. For this task, we restrict ourselves to submissions that contain exactly one URL and a title. To reduce spamming and offensive language and improve the overall quality of the data, we restricted our grounded dataset to 226 web domains and to 178 high-quality Reddit topics (i.e., "subreddits"). We also imposed constraints on turn length similar to those in place in Twitter (e.g., responses must be less than 280 characters), in order to ensure that dialogue turns are conversational and not long monologues. This filtering yielded about 3 million conversational responses and 20 million facts.[8] We split the data into train, validation and test, with the following month ranges for these different sets: years 2011-2016 for train, Jan-Mar 2017 for validation, and the rest of 2017 for test. For the test set, we selected conversational turns for which 6 or more responses were available, in order to create a multi-reference test set. Given other filtering criteria such as turn length, this yielded a 5-reference test set of size 2208 (For each instance, we set aside one of the 6 human responses to assess human performance on this task). More information about the data can be found in Qin et al. (2019), which introduced this dataset. All code and data can also be found on the DSTC Track 2 page,[9] which makes data extraction, baseline, and evaluation code available, and lets anyone recreate the training, development, validation and test sets.

3.3. Submitted Systems

The submitted systems include sequence-to-sequence models (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015) with memory network and related models (Weston et al., 2015; Sukhbaatar et al., 2015), copy-based mechanism (See et al., 2017; Gu et al., 2016; He et al., 2017), hierarchical model (Serban et al., 2016), attention mechanism (Bahdanau et al., 2015), and variational model (Kingma and Welling, 2013). The following is a brief summary of the systems based on system descriptions and private communication:

- **TeamA:** Details of this systems are unknown to us as a system description was not submitted.

- **TeamB:** It is a sequence-to-sequence model with a copying mechanism (See et al., 2017) from both the conversation history and facts. A modified beam search with some semantic clustering is proposed to discourage bland or meaningless responses.

- **TeamC:** It is a sequence-to-sequence modeling the skeleton of dialog response for pretraining, then fine-tuned with a Memory Network en-

[8]We could have easily increased the number of web domains to create a bigger dataset, but we aimed to make the task relatively accessible for participants with limited computing resources.

[9]https://github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling

coder (Sukhbaatar et al. (2015)) that utilizes retrieved top-10 related facts.

- **TeamD:** This system consists of a Memory-augmented Hierarchical Encoder-Decoder (MHRED) that extends (Serban et al., 2016), a sentence selection module to retrieve facts, and a reranker.

- **TeamF:** It is a variational generative model with a joint attention mechanism conditioning on the contexts and textual facts.

- **TeamG:** It is a variational generative model. Contexts (and response at the training stage) are encoded to extract textual fact information using an attention mechanism.

3.4. Evaluation

We evaluated response quality using both automatic and human evaluation. Since we are not considering task-oriented dialog, there is no pre-specified task and therefore no extrinsic way of measuring task success. Instead, we performed a per-response human evaluation judging each system response using crowdsourcing:

- **Relevance:** This evaluation criterion measures whether the system response is conversationally appropriate and relevant to the given K immediately preceding turns (to reduce the judges' cognitive load we set K as 2). Grounding in external sources is not involved in this judge.

- **Interest:** This evaluation criterion asks whether the produced response is interesting and informative given the document provided by the URL. To reduce cognitive load, we only considered URLs with named anchors (i.e., prefixed with '#' in the URL) and only a snippet of the document immediately following that anchor is provided to the crowdworkers. Note that models could use full web pages as input.

Both evaluation criteria were scored on a 5-point Likert scale, and finally combined the two judgments with equal weights.

In order to provide participants with preliminary results to include in their system descriptions, we also performed automatic evaluation using standard machine translation metrics, including BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and NIST (Doddington, 2002). NIST is a variant of BLEU that weights n-gram matches by their information gain, i.e., it indirectly penalizes uninformative n-grams such as "I don't" and "don't know". The final ranking of the systems was based only on human evaluation scores.

3.5. Results

3.5.1. Automatic Evaluation

The Generation Task received 26 system submissions from 7 teams. In addition to these systems, we also evaluated a "human" system (one of the six human references set aside for evaluation) and three baselines: a seq2seq baseline, a "random_human" baseline (which randomly selects human responses from the training data), and a constant baseline (which always responds "I don't know what you mean.").[10] The reason for including a constant baseline is that such a deflective response generation system can be surprisingly competitive, at least when evaluated on automatic metrics (e.g., BLEU). While the idea of such a constant baseline is relatively new, it is inspired by the idea that open-domain conversational systems trained end-to-end have a tendency to produce outputs that are relatively constant (Li et al., 2016b), such as "I don't know." The main automatic score results are shown in Table 5, and the findings for each of the metrics are as follows:

- **BLEU-4:** When evaluated on 5 references, the constant baseline, which always responds deflectively, does surprisingly well (2.87%) and outperforms all the submitted systems (ranging from 1.01% to 1.83%), and is only outperformed by humans. In further analysis, we found that reducing the number of references to one solved the problem, as almost all the systems were able to outperform the baseline according to single-reference BLEU. We suspect this deficiency of BLEU *with many references*, previously noted in Vedantam et al. (2015a), to be due to its parameterization as a precision metric. For example, if one of the gold responses happens to be "I don't know what you mean", the constant baseline gets a maximum score for that instance, irrespectively of all other references. Thus, this biases the metric towards very bland responses, as often at least one of the 5 references is somewhat deflective (e.g., contains "I don't know"). Based on these observations,

[10] This constant response was greedily selected to optimize a combination of BLEU, NIST, and METEOR on a held-out set.

we recommend to use single-reference BLEU instead of multi-reference BLEU for future DSTC tasks similar to this task, as the former gave much more meaningful results.

- **NIST-4:** The NIST score weights n-gram matches by their information gain, and effectively penalizes common n-grams such as "I don't know", which alleviates the problem with multi-reference BLEU mentioned above. None of the baselines is competitive with the top systems according to NIST-4, even when using 5 references. This suggests that NIST might be a more suitable metric than BLEU when dealing with multi-reference test sets, and it penalizes bland responses. Note that the "Random_Human" system does relatively well according to NIST-4, but this is probably due to the fact that this random baseline selects *human* sentences randomly from the training data, and human responses generally contain n-grams with more information content than machine generated n-grams.

- **METEOR:** This metric suffers from the same problem as BLEU-4, as the constant baseline performs very well on that metric and outperforms all submitted primary systems but one. We suspect this is due to the fact that METEOR (as BLEU) does not consider information gain in its scoring.

Table 5 also provides unigram and bigram diversity scores as defined in Li et al. (2016c), which are important to qualify the performance of some of the systems and baselines. Indeed, a high BLEU score (e.g., constant baseline) can be a consequence of very bland and uninformative output.

In future work, we will also consider comparing these metrics against CIDEr (Vedantam et al. (2015b)), AM-FM (D'Haro et al. (2019), Banchs et al. (2015)) Embedding Average cosine similarity, Skip-Thoughts cosine similarity, and other metrics used before in dialogue (Sharma et al. (2017)).

3.5.2. Human Evaluation

We limited evaluation to a sample of 1000 conversations and only used primary systems due to the cost of crowd-sourcing. All systems were evaluated with the same set of conversations, and results are displayed in Table 6.

Each output was judged by 3 randomly-assigned judges for Relevance and

Interest using a 5-point Likert scale. After removing spamming,[11] inter-rater agreement on a converted 3-way scale was fair, as indicated by Fleiss' Kappa at 0.39 for Relevance and 0.38 for Interest. As expected, the constant baseline performed moderately well on Relevance (2.60), but received a relatively low Interest score (constant: 2.32). The best system returned a composite score of 2.93 (Relevance: 2.99, Interest: 2.87), but is still below the human level of 3.55 (Relevance: 3.61, Interest: 3.49).

Finally, we assess the level of correlation between automatic and human scores for this task, to help determine whether it would be appropriate to rely mostly on automatic evaluation in future end-to-end response generation tasks similar to DSTC Track 2. We computed system-level correlation between overall human scores (i.e., relevance+interest) on the one hand, and each of the individual main metric on the other hand (i.e., either BLEU-4, NIST-4, and METEOR).[12] We found that automatic metrics' Spearman rank correlation coefficients (ρ) computed against human scores to be quite promising, with $\rho = 0.535$ for BLEU-4, $\rho = 0.650$ for METEOR, and $\rho = 0.669$ for NIST-4. As Table 5 suggests that BLEU-4 and NIST-4 tend to complement each other (with NIST-4 giving high scores to diverse responses, and BLEU-4 penalizing them), we also computed the correlation between the unweighted linear combination of these 3 metrics on one hand (Figure 3), and overall human scores on the other hand: this yield Spearman's $\rho = 0.754$. While this result indicates a rather strong correlation between human ratings and automatic metrics for this task, it is probably not strong enough to warrant bypassing human evaluation altogether, especially given the small sample size of this correlation analysis. Nonetheless, we consider this result to be relatively positive, as we believe it would provide participants of future end-to-end responses generation tasks a quick and relatively decent substitute to human judgment in their day-to-day (i.e., not

[11]We removed annotation of judges suspected to be spammers if their rating diverged significantly from the mean ratings of the other judges (i.e., correlation coefficient close to zero.) Such a situation is usually a sign that the judge is either rating deterministically without looking at the task (e.g., always selecting the first option in the list or ratings) or is rating randomly.

[12]Note that we computed system-level rather that sentence-level correlation, as the BLEU-4 and NIST-4 metrics were designed to be computed at a corpus rather than sentence level, as some of their underlying statistics (e.g., 4-gram matches) cannot be reliably computed on single turns or sentences.

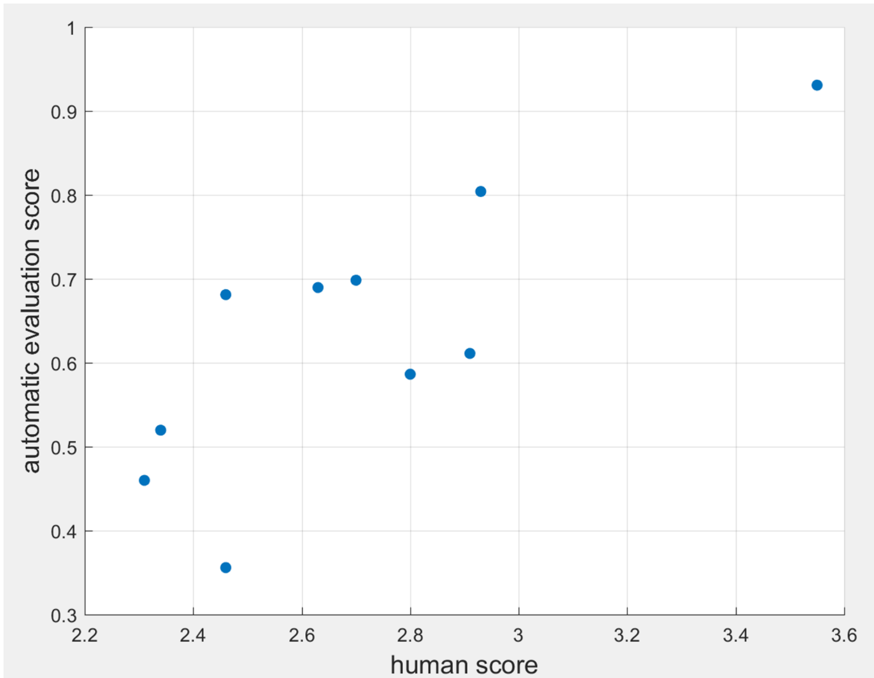

Figure 3: System-level correlation between overall human scores (relevance+interest) and automatic evaluation (unweighted linear combinatation of BLEU-4, NIST-4, and METEOR).

final) system performance evaluations.

3.6. Summary

The sentence generation task challenged participants to produce interesting and informative end-to-end conversational responses that drew on textual background knowledge. In this respect, the task was significantly more challenging that the DSTC6 task that was focused on the conversational dimensions of response generation. In general, competing system outputs were judged by humans to be more relevant and interesting than our constant and random baselines. It is also clear, however, that the quality gap between human and system responses is substantial, indicating that there is considerable space for research in future algorithmic improvements. For the future work, one line of investigation will be to explore the effect of other mechanism to extract information from the textual grounding, such as off-the-shelf machine reading models including BERT Devlin et al. (2018). Multimodal grounding is another line of future work.

4. Audio Visual Scene-aware Dialog Track

In this track, we consider a new research target: a dialog system that can discuss dynamic scenes with humans. This lies at the intersection of research in natural language processing, computer vision, and audio processing. As described above, end-to-end dialog modeling using paired input and output sentences has been proposed as a way to reduce the cost of data preparation and system development. Such end-to-end approaches have been shown to better handle flexible conversations by enabling model training on large conversational datasets (Vinyals and Le, 2015; Hori et al., 2019c). However, current dialog systems cannot understand a scene and have a conversation about what is going on in it. To develop systems that can carry on a conversation about objects and events taking place around the machines or the users, systems need to understand not only the dialog history but also the video and audio information in the scene. In the field of computer vision, interaction with humans about visual information has been explored in *visual question answering* (VQA) by Antol et al. (2015) and *Visual Dialog* by Das et al. (2017). These tasks have been the focus of intense research, aiming to (1) generate answers to questions about things and events in a single static image and (2) hold a meaningful dialog with humans about an image using natural, conversational language in an end-to-end framework. While VQA and visual dialog take significant steps towards human-machine interaction, they only consider a single static image. Most real-world scenarios, such as helping visually impaired users or intelligent home assistants, involve time-varying information. Thus, they need to be able to process video information to understanding the content and temporal dynamics of a scene. To capture the semantics of dynamic scenes, recent research has focused on *video description*. The state of the art in video description uses multimodal fusion to combine different input modalities (feature types), such as the attention-based fusion of spatio-temporal motion features and audio features proposed by Hori et al. (2017).

Since the recent revolution of neural network models allows us to combine different modules into a single end-to-end differentiable network, this framework allow us to build scene-aware dialog systems by combining end-to-end dialog and multimodal video description approaches. We can simultaneously input video features and user utterances into an encoder-decoder-based system whose outputs are natural-language responses.

To advance this goal, we introduce a new dataset of human dialogues about videos. As the subject matter of Audio Visual Scene-aware Dialog (AVSD), we used the short video clips of the Charades dataset (Sigurdsson et al., 2016): simple videos of real people performing everyday actions in real-world settings, with natural audio. The baseline system we provided incorporated technologies for video description into an end-to-end dialog system (Hori et al., 2018a). We made the dataset, code, and model publicly available for a new Audio Visual Scene-Aware Dialog (AVSD) Challenge at DSTC7.

4.1. Task definition

In this track, the system must generate responses to a user input in the context of a given dialog. The target of *VQA* and *Visual Dialog* is sentence selection based on information retrieval. For real-world application, however, spoken dialog systems cannot simply select from a small set of pre-determined sentences. Instead, they need to immediately output a response to a user input. For this reason, in this track we focus on sentence generation rather than sentence selection. In this track, the system's task is to use a dialog history (the previous rounds of questions and answers in a dialog between user and system) and (optionally) a brief video script, plus (in one version of the task) the visual and audio information from the input video, to answer a next question about the video. There are two tasks, each with two versions (a and b):

Task 1: Video and Text (a) Using the video and text training data provided but no external data sources, other than publicly available pre-trained feature extraction models (b) Also using external data for training.

Task 2: Text Only (a) Do not use the input videos nor their audio tracks for training or testing. Use only the text training data (dialog history and video script) provided. (b) Any publicly available text data may be used for training.

4.2. Data

To set up the Audio Visual Scene-Aware Dialog (AVSD) track, we collected (in Alamri et al. (2018a)) text-based dialogs about short videos from the Charades dataset (Sigurdsson et al., 2016)[13], which consists of untrimmed and multi-action videos along with a brief script for each video. The data collection paradigm for dialogs was similar to the one described by Das et al. (2016), in which for each image, two parties interacted via a text interface to yield a dialog. In Das et al. (2016), each dialog consisted of a sequence of questions and answers about an image. In our audio visual scene-aware dialog case, two parties had a discussion about events in a video. One of the two parties played the role of an answerer who had already watched the video and read the video script. The answerer answered questions asked by their counterpart, the questioner. The questioner was not allowed to watch the video but was able to see the first, middle, and last frames of the video as single static images. The two had 10 rounds of Q and A, in which the questioner asked about the events that happened in the video. At the end, the questioner summarized the events in the video as a video description.

Table 7 shows an example of a dialogue, and Table 8 shows the size of the dataset split into training, validation, and test sets. The questions and answers of the AVSD dataset mainly consist of 5 to 8 words, making them longer and more descriptive than those of VQA and Visual Dialog. Figure 4 shows the distributions of word 4-grams and average length of sentences in the questions and answers of the prototype data set of AVSD Hori et al. (2019a), compared with those of VQA and Visual Dialog (VisDial).

The dialog contains questions about objects, actions, and audio information in the videos. Although we tried to collect questions directly relevant to the event displayed, some questions refer to abstract information in the video, such as how the videos begin and the duration of the videos.

4.3. Evaluation

In this challenge, the quality of a system's automatically generated sentences is evaluated using objective measures. These determine how similar the generated responses are to groundtruth responses from humans, as well as how natural and informative the responses are. In addition to the ground truth response that was given by the answerer during dialog collection, we collected 5 additional human-generated responses for the test videos. To collect these additional responses, we provided 5 humans with all of the information that the answerer had in the original dialog: we asked them to

[13]http://allenai.org/plato/charades/

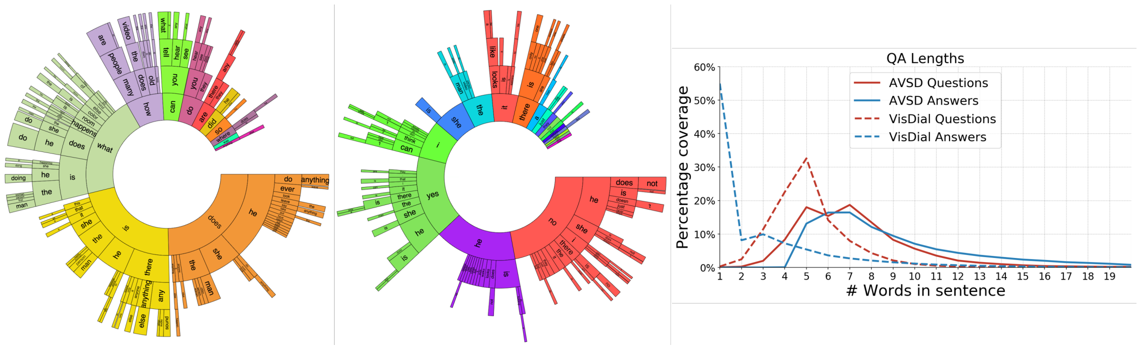

Figure 4: The distributions of word 4-grams in the questions (left) and answers (middle) of the prototype data set of the AVSD, and the average length (right) of the sentences of the VQA and the prototype data set of the AVSD. The actions were mainly asked by the questioners. There are some questions regarding audio information. Half of the answers are Yes/No. The questions and answers of AVSD are longer than those of VQA. More descriptive sentences were generated for AVSD.

answer the question after watching a video and reading the video script and the dialog history between the questioner and answerer about the video. The reason why the humans need to read the history of the dialog before answering is that there are some dependencies between each question and the the previous question/answer pairs in the sequence (Alamri et al., 2019). A typical pattern is when questions contain prepositions such as "it"— the humans cannot answer the questions if they don't know what the word "it" refers to.

We evaluated the automatically generated answers by comparing with the 6 ground truth sentences (one original answer and 5 subsequently collected answers). We used the MSCOCO evaluation tool for objective evaluation of system outputs[14]. The supported metrics include word-overlap-based metrics such as BLEU, METEOR, ROUGE_L, and CIDEr.

We also collected human ratings for each system response using a 5-point Likert Scale, where humans rated system responses given a dialog context as: 5 for very good, 4 for good, 3 for acceptable, 2 for poor, and 1 for very poor. Since the dataset contains questions and answers, we asked humans to consider correctness of the answers as well as the naturalness, informativeness, and appropriateness of the response according to the given context.

[14] https://github.com/tylin/coco-caption

4.4. Baseline System

We provided a baseline end-to-end dialog system that can generate answers in response to user questions about events in a video sequence. The baseline system is an LSTM-based encoder decoder with Naïve multimodal fusion (Alamri et al., 2018b). The architecture, which is similar to the Hierarchical Recurrent Encoder in Das et al. (2016), is based on Natural language Generation (NLG) technologies from Track2 of DSTC6 (modeling end-to-end conversation for Twitter customer service) (Hori et al., 2018b). The question, visual features, and dialog history are fed into corresponding LSTM-based encoders to build up a context embedding, and then the outputs of the encoders are fed into an LSTM-based decoder to generate an answer. The dialog history consists of encodings of QA pairs plus (optionally) an encoding of the video script. This is a simplified version of Hori et al. (2018a), in which multimodal fusion is performed without attention between modalities such as audio and video features. Figure 5 shows the architecture of the multimodal attention-based fusion. The baseline system does not have modality attention weights β. The full set of test data was used in Hori et al. (2018a), while the AVSD challenge at DSTC7 used 2,000 responses selected from the full set.

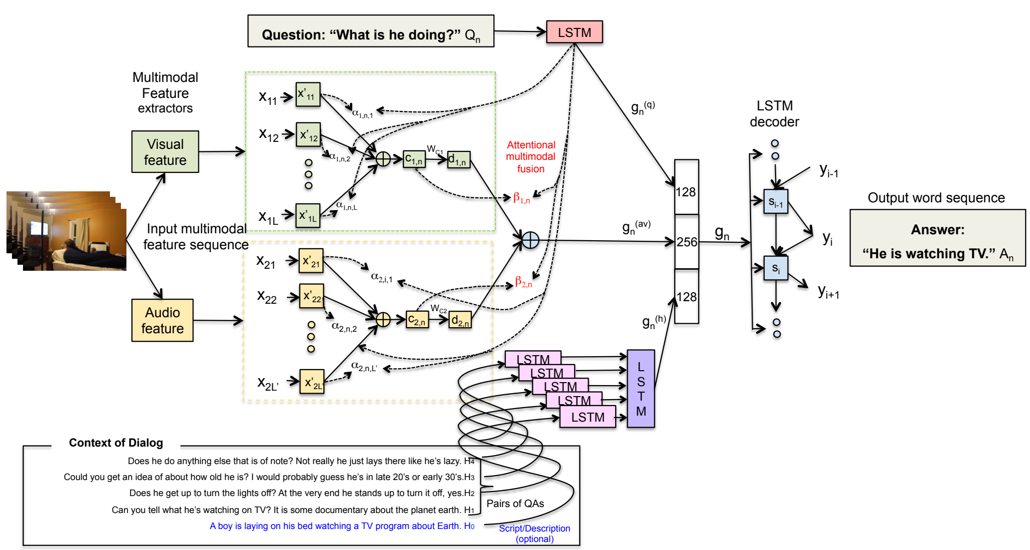

Figure 5: Attentional multimodal fusion-based video scene-aware dialog system Hori et al. (2018a)

4.5. Data Processing

4.5.1. Video Processing

We adopted the state-of-the-art I3D features Carreira and Zisserman (2017), spatiotemporal features that were developed for action recognition. The I3D model inflates the 2D filters and pooling kernels in the Inception V3 network along their temporal dimension, building 3D spatiotemporal ones. We used the output from the "Mixed_5c" layer of the I3D network to be used as video features in our framework. As a pre-processing step, we normalized all the video features to have zero mean and unit norm; the mean was computed over all the sequences in the training set for the respective feature.

In the experiments in this paper, we treated I3D-rgb (I3D features computed on a stack of 16 video frame images) and I3D-flow (I3D features computed on a stack of 16 frames of optical flow fields) as two separate modalities that are input to our multimodal attention model. To emphasize this, we refer to I3D in the results tables as I3D (rgb-flow).

4.5.2. Audio Processing

In this track, we used features extracted using a new state-of-the-art model, Audio Set VGGish (Hershey et al., 2017). Inspired by the VGG image classification architecture (Configuration A without the last group of convolutional/pooling layers), the Audio Set VGGish model operates on 0.96 sec log Mel spectrogram patches extracted from 16 kHz audio, and outputs a 128-dimensional embedding vector. The model was trained to predict an ontology of labels from only the audio tracks of millions of YouTube videos. In this work, we overlap frames of input to the VGGish network by 50%, meaning an Audio Set VGGish feature vector is output every 0.48 sec.

4.6. Submitted Systems

We received 32 sets of system outputs for the AVSD task, from 9 teams, and eight system description papers were accepted (Sanabria et al., 2019; Nguyen et al., 2019; Pasunuru and Bansal, 2019; Yeh et al., 2019; Zhuang et al., 2019; Kumar et al., 2019; Lin et al., 2019; Le et al., 2019). Table 9 shows the baseline and submitted systems with their brief specifications including Encoder-decoder Model type, Multimodal fusion type, and Additional techniques, models, and data sets. Most systems employed an LSTM, Bi-LSTM, or GRU encoder/decoder. Some systems used hierarchical and attention frameworks. Furthermore, several additional techniques were introduced to improve the response quality, such as MMI and Episodic Memory Module.

4.7. Results

The best system applied "Hierarchical Attention mechanisms to combine text and video," which was proposed in Hori et al. (2018a). Table 10 shows the evaluation results for the baseline and all systems. Figures 6–8 show the human ratings for each system in several ways. The systems are shown in the same order on the x-axis for all three figures. Figure 6 shows the mean and the standard deviation of the human ratings for each system (across all responses and all raters for that system). Figure 7 shows the distributions of the mean human rating score for each sentence for each system. Figure 8 shows the distribution of all human rating scores for each system across all sentences. In this Figure, the area for each score of the violin plot shows a count of the number of scores of each level on the Likert scale. The ratings of the reference system (labeled "Ref," at the far left of each figure) are ratings for the ground truth sentences extracted from the original QA data of the AVSD dataset. The baseline system is labeled "Base." The Reference system ("Ref") had the best human ratings: it had the highest mean rating in Fig. 6, the highest median sentence rating in Fig. 7 and the most sentences rated as level 5 ("Very good") in Fig. 8. The worst system (at the right) had a much lower mean rating and a long tail of poorly rated sentences.

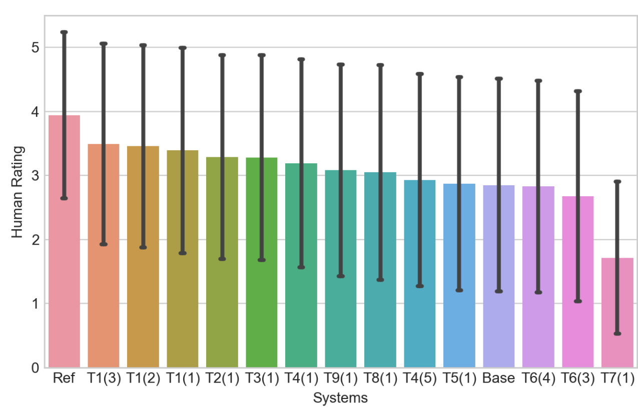

Figure 6: Mean and standard deviation of human rating score.

In Hori et al. (2018b), the reported human ratings of end-to-end conversation models for Twitter customer service data were distributed fairly

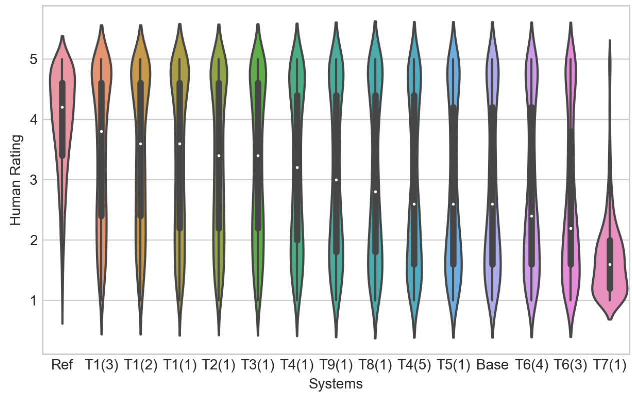

Figure 7: Distribution of human scores averaged sentence-by-sentence.

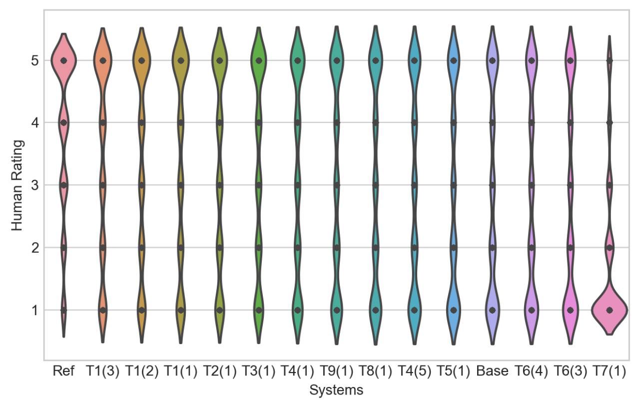

Figure 8: Distribution of human rating score for each level of scores.

smoothly in the range from 1 to 5. In contrast, the human ratings of responses in this AVSD track were more bimodal, tending to be either very low or very high (more like a binary split into "good" and "bad" answers). This is because the quality of the answers depends on the answer correctness in response to the questions, and incorrect answers result in drastically lower human rating scores. The best system generated mostly correct answers, and the worst system generated mostly incorrect answers.

4.8. Summary and Discussion

We introduced a new challenge task and dataset for Audio Visual Scene-Aware Dialog (AVSD) in DSTC7. This is the first attempt to combine end-to-end conversation and end-to-end multimodal video description models into a single end-to-end differentiable network to build scene-aware dialog systems.

The best system applied hierarchical attention mechanisms to combine text and visual information, improving by 22% over the human ratings of the baseline system. The language models trained from QA (without video or audio) are still strong approaches.

After the AVSD challenge at DSTC7, Alamri et al. (2019) reported the performance of sentence selection (as opposed to sentence generation, which was used in this AVSD challenge) using the AVSD dataset. In the paper, `Question (Q)`, `V (Video)`, `Dialog History (DH)`, and `Audio (A)` were fused. The addition of audio features generally improves model performance (`Q+V` to `Q+V+A` being the exception). Interestingly, the model performance improves even more when combined with dialog history and video features (`Q+DH+V+A`) for some metrics, indicating that audio signals still provide complementary knowledge to the video signals despite their close relationship.

Further, it is found that the best performance is achieved when including text features extracted from the available summary (video script). Surprisingly, systems that use such manual descriptions enable performance close to the best system, even without using the audio-visual features. However, such summaries are unavailable in the real world, posing challenges during deployment. Recently, Hori et al. (2019b) proposed an approach to transfer the power of the teacher model trained using summaries to a student model that does not need the summary features.

5. Conclusion and Future Directions

In this paper, we have described the seventh dialog system technology challenge (DSTC7) and the three selected tasks: sentence selection, sentence generation, and audio visual scene-aware dialog. The sentence selection track targeted the process of determining the best response given several possible answers or detecting when none candidate was suitable over two different datasets. The sentence generation track provided a testbed for knowledge-grounded response generation, with the aim of creating more controllable generators. The audio visual scene-aware dialog track raised a new problem in which dialog is generated about a given video, targeting multimodal approaches and extending the capabilities of the dialog systems to combine information from different sources.

All of the data described in this paper are provided as a large-scale benchmark of dialog systems from several viewpoints to support future dialog system research. Although submitted systems improved in all cases the baseline results, several major challenges for dialog systems still remain. For example, transferring models trained on large-scale data-sets to a variety of domains that do not have enough data is a known issue for dialog systems, as mentioned in DSTC3. Unfortunately, end-to-end systems do not address completely this issue, which would require expanding to a larger variety of domains and to consider applying transfer-learning approaches (Ruder et al. (2019)). Other problems are related with the capabilities of the dialog systems is to identify success and better managing of errors, handle task complexity in a scalable way, and the integration of multiple sources of information.

As following the raised problems in DSTC7, four tasks are proposed as the eighth edition of the dialog system technology challenge (DSTC8). Sentence selection task, track 1 in DSTC7, was extended not only a next utterance selection task but also predicting a task success and a conversation disentanglement. Audio visual scene aware dialog, track 3 in DSTC7, was also continued in the next challenge to explore a fusion between vision and dialog. Other two tasks, multi-domain task completion and scheme based dialog state tracking, were proposed as new challenges in DSTC8. Both tracks aim to build accurate task-oriented dialog systems on different approaches. Multi-domain task completion track focuses on dialog complexity and scaling to new domains as we previously focused on DSTC3. Scheme guided dialog state tracking focuses on dialog state tracking itself, even if the state space is new for the trained state tracker.

We expect to continue the challenge in the future, providing new testbeds that work towards the remaining open problems of dialog system research, while being complementary to other challenges like Alexa Prize (Khatri et al. (2018)), ConvAI (Dinan et al. (2019)), or Dialog Breakdown Detection Challenge (Higashinaka et al. (2019)).

6. Bibliography

Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., Lee, S., Parikh, D., 2019. Audio visual scene-aware dialog, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Alamri, H., Cartillier, V., Lopes, R.G., Das, A., Wang, J., Essa, I., Batra, D., Parikh, D., Cherian, A., Marks, T.K., et al., 2018a. Audio visual scene-aware dialog (avsd) challenge at dstc7. arXiv preprint arXiv:1806.00525 .

Alamri, H., Hori, C., Marks, T.K., Batra, D., Parikh, D., 2018b. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7, in: DSTC7 at AAAI2019 Workshop.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. VQA: Visual Question Answering, in: International Conference on Computer Vision (ICCV).

Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate, in: Proc. of the International Conference on Learning Representations (ICLR).

Banchs, R.E., D'Haro, L.F., Li, H., 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23, 472–482.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset, in: CVPR.

Chen, Q., Zhu, X., Ling, Z.H., Wei, S., Jiang, H., Inkpen, D., 2017. Enhanced LSTM for natural language inference, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1657–1668. URL: `http://aclweb.org/anthology/P17-1152`, doi:10.18653/v1/P17-1152.

Chen, Q.Q., Wang, W., 2019. Sequential attention-based network for noetic end-to-end response selection, in: 7th Edition of the Dialog System Technology Challenges at AAAI 2019.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M.F., Parikh, D., Batra, D., 2016. Visual dialog. CoRR abs/1611.08669. URL: `http://arxiv.org/abs/1611.08669`, `arXiv:1611.08669`.

Das, A., Kottur, S., Moura, J.M., Lee, S., Batra, D., 2017. Learning cooperative visual dialog agents with deep reinforcement learning, in: International Conference on Computer Vision (ICCV).

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

D'Haro, L.F., Banchs, R.E., Hori, C., Li, H., 2019. Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. Computer Speech & Language 55, 200–215.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., et al., 2019. The second conversational intelligence challenge (convai2). arXiv preprint arXiv:1902.00098 .

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, in: Proceedings of the Second International Conference on Human Language Technology Research, pp. 138–145.

Ganhotra, J., Patel, S.S., Fadnis, K.P., 2019. Knowledge-incorporating ESIM models for response selection in retrieval-based dialog systems, in: 7th Edition of the Dialog System Technology Challenges at AAAI 2019.

Gao, X., Lee, S., Zhang, Y., Brockett, C., Galley, M., Gao, J., Dolan, B., 2019. Jointly optimizing diversity and relevance in neural response generation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics .

Ghazvininejad, M., Brockett, C., Chang, M., Dolan, B., Gao, J., Yih, W., Galley, M., 2018. A knowledge-grounded neural conversation model. AAAI .

Gu, J., Lu, Z., Li, H., Li, V.O., 2016. Incorporating copying mechanism in sequence-to-sequence learning, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

He, S., Liu, C., Liu, K., Zhao, J., 2017. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning, in: ACL, pp. 199–208.

Henderson, M., Thomson, B., Williams, J.D., 2014a. The second dialog state tracking challenge, in: Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 263–272.

Henderson, M., Thomson, B., Williams, J.D., 2014b. The third dialog state tracking challenge, in: Spoken Language Technology Workshop (SLT), 2014 IEEE, IEEE. pp. 324–329.

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K., 2017. CNN architectures for large-scale audio classification, in: ICASSP.

Higashinaka, R., D'Haro, L.F., Shawar, B.A., Banchs, R., Funakoshi, K., Inaba, M., Tsunomori, Y., Takahashi, T., Sedoc, J., 2019. Overview of the dialogue breakdown detection challenge 4, in: 10th International Workshop on Spoken Dialog Systems (IWSDS). URL: http://workshop.colips.org/wochat/@iwsds2019/documents/dbdc4-overview-higashinak

Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T.K., Cartillier, V., Lopes, R.G., Das, A., et al., 2019a. End-to-end audio visual scene-aware dialog using multimodal attention-based video features, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 2352–2356.

Hori, C., Alamri, H., Wang, J., Winchern, G., Hori, T., Cherian, A., Marks, T.K., Cartillier, V., Lopes, R.G., Das, A., et al., 2018a. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. arXiv preprint arXiv:1806.08409 .

Hori, C., Hori, T., 2017a. End-to-end conversation modeling track in DSTC6, in: Dialog System Technology Challenges 6. URL: http://workshop.colips.org/dstc6/papers/track2_overview_hori.pdf.

Hori, C., Hori, T., 2017b. End-to-end conversation modeling track in DSTC6. arXiv:1706.07440 .

Hori, C., Hori, T., Cherian, A., Marks, T.K., 2019b. Joint student-teacher learning for audio-visual scene-aware dialog, in: Interspeech2019, ISCA.

Hori, C., Hori, T., Lee, T.Y., Zhang, Z., Harsham, B., Hershey, J.R., Marks, T.K., Sumi, K., 2017. Attention-based multimodal fusion for video description, in: ICCV.

Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., Kim, S., 2019c. Overview of the sixth dialog system technology challenge: Dstc6. Computer Speech & Language 55, 1–25.

Hori, C., Perez, J., Higasinaka, R., Hori, T., Boureau, Y.L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., Kim, S., 2018b. Overview of the sixth dialog system technology challenge: DSTC6. Computer Speech and Language Special issue on DSTC6.

Jiang, Y., Kummerfeld, J.K., Lasecki, W.S., 2017. Understanding task design trade-offs in crowdsourced paraphrase collection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). URL: http://aclweb.org/anthology/P17-2017.

Khatri, C., Hedayatnia, B., Venkatesh, A., Nunn, J., Pan, Y., Liu, Q., Song, H., Gottardi, A., Kwatra, S., Pancholi, S., et al., 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. arXiv preprint arXiv:1812.10757 .

Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J.D., Henderson, M., 2017. The fourth dialog state tracking challenge, in: Dialogues with Social Robots. Springer, pp. 435–449.

Kim, S., D'Haro, L.F., Banchs, R.E., Williams, J.D., Henderson, M., Yoshino, K., 2016. The fifth dialog state tracking challenge, in: Spoken Language Technology Workshop (SLT), 2016 IEEEover, IEEE. pp. 511–517.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 .

Kumar, S.H., Okur, E., Sahay, S., Leanos, J.J.A., Huang, J., Nachman, L., 2019. Context, attention and audio feature explorations for audio visual scene-aware dialoge, in: DSTC7 at AAAI2019 workshop.

Kummerfeld, J.K., 2019. Slate: A super-lightweight annotation tool for experts, in: Proceedings of ACL 2019, System Demonstrations.

Kummerfeld, J.K., Gouravajhala, S.R., Peper, J., Athreya, V., Gunasekara, C., Ganhotra, J., Patel, S.S., Polymenakos, L., Lasecki, W.S., 2018. Analyzing assumptions in conversation disentanglement research through the lens of a new dataset and model. ArXiv e-prints URL: https://arxiv.org/pdf/1810.11118.pdf, arXiv:1810.11118.

Lavie, A., Agarwal, A., 2007. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments, in: Proc. of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 228–231. URL: http://dl.acm.org/citation.cfm?id=1626355.1626389.

Le, H., Hoi, S., Sahoo, D., Chen, N., 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features, in: DSTC7 at AAAI2019 workshop.

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2016a. A diversity-promoting objective function for neural conversation models, in: NAACL-HLT.

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2016b. A diversity-promoting objective function for neural conversation models, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 110–119.

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B., 2016c. A diversity-promoting objective function for neural conversation models. NAACL-HLT .

Lin, K.Y., Hsu, C.C., Chen, Y.N., Ku, L.W., 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation, in: DSTC7 at AAAI2019 workshop.

Lowe, R., Pow, N., Serban, I., Pineau, J., 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, in: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Prague, Czech Republic. pp. 285–294. URL: http://aclweb.org/anthology/W15-4640.

Nguyen, D., Sharma, S., Schulz, H., Asri, L.E., 2019. From film to video: Multi-turn question answering with multi-modal context, in: DSTC7 at AAAI2019 workshop.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation. ACL .

Pasunuru, R.R., Bansal, M., 2019. Dstc7-avsd: Scene-aware video-dialogue systems with dual attention, in: DSTC7 at AAAI2019 workshop.

Perez, J., Boureau, Y.L., Bordes, A., 2017. Dialog system technology challenge 6 overview of track 1 - end-to-end goal-oriented dialog learning, in: Dialog System Technology Challenges 6. URL: http://workshop.colips.org/dstc6/papers/track1_overview_perez.pdf.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 2227–2237. URL: http://aclweb.org/anthology/N18-1202, doi:10.18653/v1/N18-1202.

Qin, L., Galley, M., Brockett, C., Liu, X., Gao, X., Dolan, B., Choi, Y., Gao, J., 2019. Conversing by reading: Contentful neural conversation with on-demand machine reading, in: Proc. of ACL.

Ritter, A., Cherry, C., Dolan, W.B., 2011. Data-driven response generation in social media. EMNLP .

Ruder, S., Peters, M.E., Swayamdipta, S., Wolf, T., 2019. Transfer learning in natural language processing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pp. 15–18.

Sanabria, R., Palaskar, S., Metze, F., 2019. Cmu sinbad submission for the dstc7 avsd challenge, in: DSTC7 at AAAI2019 workshop.

See, A., Liu, P.J., Manning, C.D., 2017. Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1073–1083. doi:10.18653/v1/P17-1099.

Serban, I.V., Lowe, R., Henderson, P., Charlin, L., Pineau, J., 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. Dialogue & Discourse 9, 1–49. URL: http://dad.uni-bielefeld.de/index.php/dad/article/view/3690, doi:10.5087/dad.2018.101.

Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J., 2016. Building end-to-end dialogue systems using generative hierarchical neural network models, in: Proc. of AAAI.

Shang, L., Lu, Z., Li, H., 2015. Neural responding machine for short-text conversation. ACL-IJCNLP .

Sharma, S., El Asri, L., Schulz, H., Zumer, J., 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. CoRR abs/1706.09799. URL: http://arxiv.org/abs/1706.09799.

Sigurdsson, G.A., Varol, G., Wang, X., Laptev, I., Farhadi, A., Gupta, A., 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. ArXiv URL: http://arxiv.org/abs/1604.01753, arXiv:1604.01753.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.Y., Gao, J., Dolan, B., 2015. A neural network approach to context-sensitive generation of conversational responses. NAACL-HLT .

Sukhbaatar, S., szlam, a., Weston, J., Fergus, R., 2015. End-to-end memory networks, in: Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 2440–2448.

Vedantam, R., Zitnick, C.L., Parikh, D., 2015a. CIDEr: Consensus-based image description evaluation., in: CVPR, pp. 4566–4575.

Vedantam, R., Zitnick, C.L., Parikh, D., 2015b. CIDEr: Consensus-based image description evaluation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 4566–4575.

Vinyals, O., Le, Q., 2015. A neural conversational model. ICML .

Weston, J., Chopra, S., Bordes, A., 2015. Memory networks. ICLR .

Williams, J., Raux, A., Ramachandran, D., Black, A., 2013. The dialog state tracking challenge, in: Proceedings of the SIGDIAL 2013 Conference, pp. 404–413.

Yeh, Y.T., Lin, T.C., Cheng, H.H., Deng, Y.H., Su, S.Y., Chen, Y.N., 2019. Reactive multi-stage feature fusion for multimodal dialogue modeling, in: DSTC7 at AAAI2019 workshop.

Zhuang, B., Wang, W., Shinozaki, T., 2019. Investigation of attention-based multimodal fusion and maximum mutual information objective for dstc7 track3, in: DSTC7 at AAAI2019 workshop.

Team	Model Type	External Data Use	Used Raw Advising	Val in Train	Model Details
1	CNN	-	No	Yes	Combination of CNN for utterance representation and GRU for modeling the dialogue.
2	LSTM	-	Yes	No	ESIM with an aggregation scheme to capture dialog-specific aspects of the data + ELMo.
3	LSTM	Embeddings	Yes	No	ESIM + a filtering stage for subtask 2.
4	LSTM	-	No	No	ESIM with (1) enhanced word embeddings to address OOV issues, (2) an attentive hierarchical recurrent encoder, and (3) an additional layer before the softmax.
6	Ensemble	-	No	No	An ensemble of CNNs.
7	LSTM	-	No	Yes	LSTM representation of utterances followed by a convolutional layer.
8	Other	-	Yes	No	A multi-level retrieval-based approach that aggregates similarity measures between the context and the candidate response on the sequence and word levels.
10	LSTM	TF-IDF Extraction	No	No	ESIM with matching against similar dialogues in training, and an extra filtering step for subtask 2.
12	RNN	TF-IDF Extraction	No	No	BoW over ELMo with context as an RNN.
13	Ensemble	Embeddings	No	No	Ensemble approach, combining a Dynamic-Pooling LSTM, a Recurrent Transformer and a Hierarchical LSTM.
14	Ensemble	-	No	No	An ensemble using voting, combining the baseline LSTM, a GRU variant, Doc2Vec, TF-IDF, and LSI.
15	Memory	Memory	No	No	Memory network with an LSTM cell.
16	LSTM	-	No	No	ESIM with utterance-level attention, plus additional features.
17	Memory	Memory & Embeddings	Yes	No	Self-attentive memory network, with external advising data in memory and external ubuntu data for embedding training.
18	GRU	-	No	No	Stacked Bi-GRU network with attention, aggreagting attention across the temporal dimension followed by a CNN and softmax.
19	LSTM	-	No	Yes	Bidirectional LSTM memory network.
20	CNN	-	No	Yes	CNN with attention and a pointer network, plus a novel top-k attention mechanism.

Table 2: Summary of approaches used by participants for track-1. All teams applied neural approaches, with ESIM being a popular basis for system development. External data refers to the man pages for Ubuntu, and course information for Advising. Raw advising refers to the variant of the training data in which the complete dialogues and paraphrase sets are provided. Teams 5, 9 and 11 did not provide descriptions of their approaches. For further details, see the system description papers presented at the DSTC workshop.

	Ubuntu, Subtask				Advising, Subtask			
Team	1	2	4	5	1	3	4	5
3	**0.819**	0.145	**0.842**	**0.822**	**0.485**	**0.592**	**0.537**	**0.485**
4	0.772	-	-	-	0.451	-	-	-
17	0.705	-	-	0.722	0.434	-	-	0.461
13	0.729	-	0.736	0.635	0.458	0.461	0.474	0.390
2	0.672	0.033	0.713	0.672	0.430	0.540	0.479	0.430
10	0.651	**0.307**	0.696	0.693	0.361	0.434	0.262	0.361
18	0.690	0.000	0.721	0.710	0.287	0.380	0.398	0.326
8	0.641	-	0.527	-	0.310	0.433	0.233	-
16	0.629	0.000	0.683	-	0.280	-	0.370	-
15	0.473	-	-	0.478	0.300	-	-	0.236
7	0.525	-	0.411	-	-	-	-	-
11	-	-	-	-	0.075	0.232	-	-
12	0.077	-	0.000	0.077	0.075	0.232	0.000	0.075
1	0.580	-	-	-	0.239	-	-	-
6	-	-	-	-	0.245	-	-	-
9	0.482	-	-	-	-	-	-	-
14	0.008	-	0.072	-	-	-	-	-
19	0.265	-	-	-	0.180	-	-	-
5	0.076	-	-	-	-	-	-	-
20	0.002	-	-	-	0.004	-	-	-

Table 3: Track-1 results, ordered by the average rank of each team across the sub-tasks they participated in. The top result in each column is in bold. For these results the metric is the average of MRR and Recall@10.

Web page info	[...] she holds the guinness world record for **surviving** the highest fall without a parachute : **10,160 metres** (**33,330 ft**) . [...] **four years later** , peter hornung-andersen and pavel theiner , two prague-based journalists , claimed that flight 367 had been mistaken for an enemy aircraft and shot down by the czechoslovak air force at an altitude of **800 metres** (2,600 ft) [...]
Turn 1	today i learned a woman fell **30,000 feet** from an airplane and **survived** [URL] .
Turn 2	the page states that a **2009 report** found the plane only fell **several hundred meters** .
Turn 3	well if she only fell a **few hundred meters** and survived then i 'm not impressed at all .
Turn 4	still pretty incredible , but quite a bit different that **10,000 meters** .

Table 4: Sample of the DSTC7 Sentence Generation data, which combines Reddit data (Turns 1-4) along with documents (extracted from Common Crawl) discussed in the conversations. The web page info was truncated for this figure to fit in a relatively small space. The **emphasis** was added by us. The [URL] links to the web page above.

System	NIST N-2	NIST N-4	BLEU(%) B-2	BLEU(%) B-4	METEOR	Diversity D-1	Diversity D-2	Avg. len
Baselines:								
Constant	0.18	0.18	12.8	2.9	7.5	0.1	0.1	8.0
Random_Human	1.63	1.64	6.7	0.9	5.9	16.0	64.7	19.2
Seq2Seq	0.91	0.92	14.8	1.8	7.0	1.4	4.8	10.6
TeamA	0.75	0.75	11.8	1.5	5.6	9.6	27.6	10.5
TeamA-c1	0.83	0.83	11.5	1.4	5.7	12.2	30.2	10.9
TeamA-c2	1.12	1.12	9.5	0.8	5.5	9.7	31.9	12.0
TeamB	2.51	**2.5**	14.4	1.8	**8.1**	10.9	32.5	15.1
TeamB-c1	1.76	1.77	13.7	**1.9**	7.6	9.4	26.7	12.8
TeamC	1.51	1.51	10.9	1.3	6.4	5.3	17.1	12.7
TeamC-c1	2.11	2.12	9.9	1.3	6.8	3.8	12.4	16.4
TeamC-c2	1.19	1.20	11.6	1.7	6.2	5.5	16.9	11.7
TeamC-c3	1.73	1.74	8.8	1.2	5.9	3.9	12.2	14.9
TeamC-c4	1.53	1.54	11.5	1.8	6.5	5.6	18.0	12.7
TeamD	2.04	2.05	11.3	1.4	6.7	9.4	33.4	14.4
TeamD-c1	0.02	0.02	6.7	0.3	3.9	2.6	16.1	6.2
TeamD-c2	0.73	0.73	9.3	0.6	5.7	4.9	31.3	10.4
TeamD-c3	0.77	0.77	9.2	0.7	5.6	4.9	30.9	10.5
TeamD-c4	0.55	0.56	8.8	0.8	5.2	6.9	35.2	9.8
TeamD-c5	1.80	1.80	10.7	0.9	6.5	5.8	29.2	13.5
TeamD-c6	1.74	1.75	12.5	1.1	6.7	5.1	20.7	13.1
TeamE	1.51	1.51	10.9	1.3	6.4	5.3	17.1	12.7
TeamE-c1	2.11	2.12	9.9	1.3	6.8	3.8	12.4	16.4
TeamE-c2	1.81	1.82	11.0	1.6	6.5	5.0	15.6	14.0
TeamE-c3	1.92	1.93	10.9	1.5	6.7	4.6	15.2	14.3
TeamF	0.01	0.01	10.2	1.0	4.6	6.4	17.6	5.4
TeamF-c1	0.01	0.01	9.0	1.3	4.1	2.4	7.2	5.1
TeamF-c2	0.04	0.04	11.2	1.4	5.0	8.4	22.4	6.3
TeamG	2.31	2.32	10.6	1.2	7.2	3.4	26.5	16.6
TeamG-c1	2.03	2.04	8.2	1.1	7.5	10.8	44.9	22.3
Human	2.62	2.65	12.4	3.1	8.3	16.7	67.0	18.8

Table 5: Automatic evaluation results for track-2. Participants submitted primary and contrastive systems, the latter being identified with a -cX suffix in their names. The primary systems (TeamA, TeamB, ...) were the ones selected by the participants for human evaluation (Table 6).

System	Relevance		Interest		Overall	
	Mean	95% CI	Mean	95% CI	Mean	95 % CI
Baselines:						
Constant	2.60	(2.560, 2.644)	2.32	(2.281, 2.364)	2.46	(2.424, 2.500)
Random	2.32	(2.269, 2.371)	2.35	(2.303, 2.401)	2.34	(2.288, 2.384)
Seq2Seq	2.91	(2.858, 2.963)	2.68	(2.632, 2.730)	2.80	(2.748, 2.844)
TeamA	2.32	(2.267, 2.368)	2.30	(2.252, 2.351)	2.31	(2.262, 2.358)
TeamB	2.99	(2.938, 3.042)	**2.87**	(2.822, 2.922)	**2.93**	(2.882, 2.979)
TeamC	**3.05**	(3.009, 3.093)	2.77	(2.735, 2.812)	2.91	(2.875, 2.950)
TeamD	2.69	(2.635, 2.743)	2.58	(2.527, 2.632)	2.63	(2.583, 2.685)
TeamF	2.52	(2.461, 2.572)	2.40	(2.352, 2.457)	2.46	(2.409, 2.512)
TeamG	2.82	(2.771, 2.870)	2.57	(2.525, 2.619)	2.70	(2.650, 2.742)
Human	3.61	(3.554, 3.658)	3.49	(3.434, 3.539)	3.55	(3.497, 3.596)

Table 6: Human evaluation results for track-2. The systems evaluated here are the same as the primary systems in Table 5. Note that we do not report the results of TeamE as their primary system was identical to TeamC's (due to miss-communication at submission time). The best system according to human evaluation (TeamB) also obtained the best NIST-4 and METEOR scores.

	Questioner	Answerer
QA1	What kind of room does this appear to be?	He appears to be in the bedroom.
QA2	How does the video begin?	By him entering the room.
QA3	Does he have anything in his hands?	He pick up a towel and folds it.
QA4	What does he do with it ?	He just folds them and leaves them on the chair.
QA5	What does he do next?	Nothing much except this activity.
QA6	Does he speak in the video?	No he did not speak at all.
QA7	Is there anyone else in room at all?	No he appears alone there.
QA8	Can you see or hear any pets in the video?	No pets to see in this clip.
QA9	Is there any noise in the video of importance?	Not any noise important there.
QA10	Are there any other actions in the video?	Nothing else important to know.

Table 7: An example dialog from the AVSD dataset.

	training	validation	test
# of dialogs	7,659	1,787	1,710
# of turns	153,180	35,740	13,490
# of words	1,450,754	339,006	110,252

Table 8: The dialog data for the DSTC7 AVSD track. The test videos for this challenge were selected from the official test data of the Charades challenge.

Team	Encoder-decorder type	Multimodal fusion type	Additional techniques/data
baseline	LSTM	Naïve fusion	
team_1	Bidirectional Gated Recurrent Units (GRU) based encode, Conditional Gated Recurrent Units (CGRU) based decoder	Hierarchical attention	ResNeXt, Transfer learning using How2 dataset
team_2	FiLM Attention Hierarchical Recurrent Encoder Decoder (FA-HRED), LSTM	Naïve fusion	FiLM
team_3	Dual attention LSTM encoder,	Cross-attention fusion	Similarity matrix
team_4	LSTM/GRU encoder, Top-down Attention LSTM/GRU decoder	Muti-stage fusion, 1x1 Convolution fusion, Multi-head Attention	
team_5	Bi-LSTM and LSTM encoder, LSTM decoder	Attentional multimodal fusion	MMI objective
team_6	LSTM encoder-decoder	Attentional multimodal fusion	Topic-base Conceptual model, ConvNet, AclMet
team_7	–	–	–
team_8	Bi-LSTM/LSTM encoder, Attention-based GRU encoder, LSTM decoder	Entropy-enhanced Dynamic Memory Network (DMN)	Episodic Memory Module
team_9	GRU encoder-decoder	Question-to-Caption/Multimodal attention	

$^{+}$Team 7 did not submit a system description paper to the DSTC7 workshop.

Table 9: Submitted systems to the AVSD Track.

Team	Entry	text only	video	caption and/or summary	extra	prototype	Bleu_4	METEOR	ROUGE_L	CIDEr	Human rating
Team 1	(1)	✓		✓	✓		0.376	0.264	0.554	1.076	3.394
	(2)		✓	✓	✓		0.387	0.266	**0.564**	1.087	3.459
	(3)		✓	✓			**0.394**	**0.267**	0.563	**1.094**	**3.491**
	(4)	✓		✓			0.364	0.254	0.543	1.006	-
Team 2	(1)		✓	✓			0.360	0.249	0.544	0.997	3.288
	(2)	✓	✓	✓			0.323	0.231	0.510	0.843	
	(3)	✓		✓			0.343	0.243	0.536	0.920	
	(4)	✓		✓			0.340	0.228	0.518	0.851	
	(5)			✓		✓	0.349	0.242	0.536	0.947	
	(6)		✓	✓		✓	0.316	0.224	0.505	0.795	
	(7)		✓	✓		✓	0.319	0.228	0.513	0.836	
	(8)	✓		✓		✓	0.323	0.220	0.501	0.799	
Team 3	(1)		✓	✓			0.337	0.242	0.532	0.957	3.279
Team 4	(1)		✓	✓		✓	0.342	0.223	0.504	0.837	3.188
	(2)		✓	✓			0.345	0.224	0.505	0.877	
	(3)		✓	✓		✓	0.342	0.223	0.504	0.836	
	(4)	✓		✓			0.304	0.207	0.477	0.731	
	(5)	✓		✓			0.304	0.206	0.475	0.729	2.928
Team 5	(1)		✓			✓	0.293	0.221	0.486	0.761	2.869
	(2)		✓			✓	0.302	0.222	0.488	0.770	
	(3)		✓			✓	0.302	0.222	0.487	0.769	
	(4)		✓			✓	0.296	0.219	0.484	0.745	
	(5)		✓			✓	0.283	0.217	0.480	0.731	
Team 6	(1)	✓	✓	✓		✓	0.307	0.213	0.469	0.701	
	(2)	✓	✓	✓		✓	0.307	0.215	0.479	0.733	
	(3)	✓	✓	✓		✓	0.278	0.198	0.442	0.614	2.675
	(4)	✓	✓	✓		✓	0.310	0.217	0.483	0.718	2.827
Team 7	(1)	✓		✓			0.056	0.096	0.236	0.085	1.715
Team 8	(1)		✓	✓			0.310	0.241	0.527	0.912	3.048
	(2)		✓	✓			0.307	0.239	0.525	0.915	
Team 9	(1)	✓		✓			0.310	0.242	0.515	0.856	3.080
	(2)		✓				0.315	0.239	0.509	0.848	
Reference											3.938
Baseline w/o audio			✓				0.305	0.217	0.481	0.733	
Baseline			✓				0.309	0.215	0.487	0.746	2.848

Table 10: Evaluation results with word-overlapping-based objective measures based on 6 references and a subjective measure based on 5-level ratings for the AVSD track. Under this evaluation, the human rating for the original answers was **3.938**.