Storytelling with Dialogue: A *Critical Role* Dungeons and Dragons Dataset

Revanth Rameshkumar
Microsoft
reramesh@microsoft.com

Peter Bailey
Microsoft
pbailey@microsoft.com

Abstract

This paper describes the *Critical Role* Dungeons and Dragons Dataset (CRD3) and related analyses. *Critical Role* is an unscripted, live-streamed show where a fixed group of people play Dungeons and Dragons, an open-ended role-playing game. The dataset is collected from 159 *Critical Role* episodes transcribed to text dialogues, consisting of 398,682 turns. It also includes corresponding abstractive summaries collected from the Fandom wiki. The dataset is linguistically unique in that the narratives are generated entirely through player collaboration and spoken interaction. For each dialogue, there are a large number of turns, multiple abstractive summaries with varying levels of detail, and semantic ties to the previous dialogues. In addition, we provide a data augmentation method that produces 34,243 summary-dialogue chunk pairs to support current neural ML approaches, and we provide an abstractive summarization benchmark and evaluation.

1 Introduction

Artificial intelligence applied to human conversation remains an incredibly challenging task in computer science. Task-oriented dialogues, which are more narrowly scoped and information dense than conversational dialogue, have been the focus of recent progress in dialogue understanding (Budzianowski et al., 2018). A difficulty for hypothesis testing on non-task oriented dialogues is a lack of large datasets that are fully representative of the spontaneity and noise of real world conversation, especially in the areas of storytelling and narrative beyond long-form text or monologue. Many potential dialogue processing tasks involve multi-speaker dialogues where narrative elements are conveyed through interaction between two or more speakers. These narrative elements can include changes in the states of narrative objects, descriptions of events, or changes in the states of speakers themselves. Some explored sub-tasks for narrative understanding are topic understanding, character state tracking, and abstractive summarization. Though progress has been made in these areas, it has been on datasets where conversation has been constrained to specific topics, constrained by

	Sample Dialogue Chunk
0	TRAVIS: “i felt like i almost died and i had n’t taken care of any of the shit that got me here in the first place . i was so worried about trying to learn about these new abilities that – i felt like i got distracted . i have people i want to find and things i want to remedy .”
1	MARIHSA: “yeah . how did jester do ? no offense , but she seems like she ’s a little bit more willfully stronger than you are .”
2	TRAVIS: “i mean , fuck , it ’s really disturbing . like , she came out of there like a little kettle of popcorn , just no problem . i mean – can i see jester ? is she nearby ?”
3	MATT: “jester , are you nearby ?”
4	LAURA: “i ’m across the bar just fucking dancing alone . -lrb- laughter -rrb- .”
5	LIAM: “just sixteen candles-ing it .”
6	MARIHSA: “yep .”
7	TRAVIS: “i was worried . there were really dark times . i would hear jester singing to herself at night and then she ’d change lyrics , and then my name would be in the lyrics sometimes . every morning , she would try and cheer everybody up that was around her , but she had the muffle ? so i could n’t tell if my brain was playing tricks on me , or if she was just – i do n’t think there ’s much that gets her down . it ’s kind of inspiring .”
	Aligned Summary Chunk
0	“beau asks about jester .”
1	“fjord says he is amazed but disturbed at how well jester seems to be doing .”
2	“he says jester would try to cheer everyone up and sing , even though her mouth was gagged most of the time .”
3	“he looks over to see jester dancing alone by the end of the bar .”

Figure 1: A tokenized dialogue chunk and the associated human written summary chunk after the text alignment process. Jester, Beau, and Fjord are the aliases for Laura, Marisha, and Travis respectively.

medium of communication, or scripted (in the case of television or movies) (Forchini, 2009). With datasets that involve naturally occurring dialogue, the small amount of data per narrative or speaker makes modeling challenging.

1.1 *Critical Role* Episodes and Wiki

The *Critical Role* show[1] is a weekly unscripted, live-stream of a fixed group of people playing Dungeons and Dragons, a popular role-playing game. *Critical Role* is set in a fictional world created by the Dungeon Master (DM) Matthew Mercer.

Separate from Matthew, there are eight other players who participate in his world as role-played characters; whose actions in the game influence the fictional world (as per the DM) along with their own character's state. There are multiple objectives to the game, both hidden and explicitly stated by both parties. For example, the DM might explicitly state a quest for the players to complete or a player's character might have an explicit personal goal that needs to be met. Examples of implicit objectives are non-player characters objectives created by the DM, and a player's character's backstory that influence their actions. This definition and expansion of the fictional world, the interaction with the world, and the development of the narrative is done entirely through unscripted spoken dialogue between the DM and the other players.

Fans have maintained dialogue transcriptions for each episode as well as an online knowledge base (the Fandom wiki[2]) where details about the players, characters, world, and game sessions are continuously added to. By extracting dialogues from the *Critical Role* transcripts, CRD3 aims to provide the community with a narrative-centered dataset that is unscripted, noisy, and spontaneous; while being coherent, consistent in latent speaker attributes and personalities, and considerably longer in dialogue length than similar conversational dialogue datasets. From the wiki, we obtain human-authored, structured summaries for each episode that support tasks of narrative understanding and extraction, topic understanding and segmentation, and summarization from conversational dialogue.

1.2 Contributions

We make five contributions in this paper. First, we produce a cleaned and structured dialogue dataset extracted from the *Critical Role* transcripts (CRD3-Dialogues)[3]. Second, we provide corresponding structured abstractive summaries for each episode, mined from the Fandom wiki (CRD3-Summaries). Third, we analyze the dataset and compare it to similar datasets. Fourth, we describe our method of data augmentation via text alignment to make this data scale-appropriate for neural ML approaches, and provide these summary-dialogue chunk pairs (CRD3-SD-pairs). Finally, we construct an abstractive summarization baseline from these pairs and discuss its evaluation (CRD3-Baseline).

We believe that better abstractive summarization tools to distill information is essential given the ongoing growth of unscripted, multi-person dialogues in entertainment and business scenarios. We hope that CRD3 will support research and development for such tools.

2 Related Work

The *Critical Role* Dungeons and Dragons Dataset is a combination of story-telling dialogues structured around the game-play of Dungeons and Dragons and corresponding abstractive summarizations for each dialogue. As such, it can be compared to existing dialogue datasets and summarization datasets.

2.1 Dialogue Datasets

There are currently many existing dialogue datasets (disregarding machine-to-machine) that can be roughly grouped into task-oriented, conversational, scripted, constrained, and spontaneous dialogues (Serban et al., 2015). Task-oriented datasets address specific tasks and are constrained by an ontology (Budzianowski et al., 2018). If the task is sufficiently constrained, even a human-to-human task-oriented dialogue can lack spontaneity and noise of open domain conversation (Haber et al., 2019), (Vaidyanathan et al., 2018), (Lison and Tiedemann, 2016). Agents trained on such datasets cannot be expected to model spontaneous conversational dialogue. Scripted dialogue datasets are closer to conversational dialogue. Popular scripted dialogues come from TV shows, movies, and novels; sometimes featuring further annotations (Poria et al., 2019a), (Lison and Tiedemann, 2016), (Banchs, 2012). Though the lack of noise can be helpful in training a dialogue system, they do contain artificialities in their linguistic properties (Forchini, 2009). With datasets that do have

[1] critrole.com
[2] criticalrole.fandom.com
[3] github.com/RevanthRameshkumar/CRD3

natural conversation, either with provided topics (Rashkin et al., 2019), (Godfrey et al., 1992), (Carletta et al., 2006) or truly naturally occurring (Ritter et al., 2010),(Schrading et al., 2015), (Li et al., 2017), (Leech, 1992), (Misra et al., 2015), the larger scope and noise along with the small amount of data for individual domains, latent speaker attributes, and linguistic attributes make tasks like response generation, abstractive summarization, and speaker personality modeling more difficult (Vinyals and Le, 2015), (Black et al., 2011), (Stent et al., 2005), (Poria et al., 2019b). Story-telling and game-playing dialogues can have properties from both task-oriented and conversational dialogues, as they have specific topics or tasks and are primarily human-to-human (Gratch et al., 2007), (Hung and Chittaranjan, 2009), (Afantenos et al., 2012), (Djalali et al., 2012), (Hu et al., 2016). In story-telling dialogues there is a clear topic constraint and purpose of conveying narratives. In game-play dialogues, there are clear tasks that the speakers try to complete, to either win or progress the game. This helps reduce topic noise and increase information density, but retains natural noise like disfluencies, false starts, fragments, and spontaneity.

CRD3 has extensive storytelling and narrative building through dialogue, as well as game-playing since Dungeons and Dragons is the show's focus. The episodes are unscripted and live-streamed, so the dialogue is naturally occurring and contains a large amount of context-switching and chit-chat. Since it is spoken then transcribed to text, there exists linguistic noise as usually present in naturally spoken dialogue. Finally, the large amount of turns combined with consistent cast and persistent environments make modelling based on latent speaker and linguistic attributes more feasible.

2.2 Abstractive Summarization Datasets

Most of the recent abstractive summarization research is conducted on document datasets (news, scientific papers, and patents) (Hermann et al., 2015), (Cohan et al., 2018), (Sharma et al., 2019). However, the methods used to perform well in these domains are less effective in dialogue (movies, personal-interviews, multi-person dialogues, etc) (Kedzie et al., 2018). As (Narayan et al., 2018) noted, many of the current summarization datasets highly reward extractive approaches due to the large amount of phrasal overlap in document and summary. Dialogue summarization is underexplored in datasets. For abstractive summarization, the most popular spoken dialogue datasets are AMI and Switchboard. Others exist, but are more constrained or purely textual, (Zhou et al., 2018), (Gella et al., 2018), (Misra et al., 2015), (Louis and Sutton, 2018), (Pan et al., 2018). Notably, (Gorinski and Lapata, 2015), (Gorinski and Lapata, 2018) combine movie scripts with Wikipedia plot summaries and other metadata. Though this brings us closer to longer form abstractive dialogue summarization data, there is significant information about the plot conveyed through script notes and descriptions, and not spoken dialogue.

3 Data Collection and Preprocessing

3.1 Dungeons and Dragons

Briefly, Dungeons and Dragons is a popular role-playing game that is driven by structured story-telling. Players create characters to participate in a fictional world created by the Dungeon Master (DM). They interact with the world entirely through dialogue with the DM and use dice rolls as a way to introduce randomness to the consequences of their actions. Actions can include exploring the environment, talking to fictional characters (role played by the DM), battle, and puzzle solving.[4]

3.2 *Critical Role* Video Stream Transcripts

The CRD3 dataset consists of 159 episodes (dialogues) from two campaigns. Campaign 1 has 113 episodes and Campaign 2 has 46 episodes, with new episodes being actively added. The episodes are unscripted and live-streamed, then archived and transcribed; they are usually several hours long. Detailed episode information can be found on the Fandom wiki[5]. The episodes usually start with some out-of-narrative logistics, then proceed to the actual D&D game where the players communicate character action by in-character role-playing or by describing the characters' actions in third person. There is also substantial out of narrative chit-chat and context switching.

For each episode, we extract the names and turns from the dialogue transcript and clean the data as much as possible. We try to resolve the inconsistencies in spelling of speaker names, use of quotes, onomatopoeia, speaker aliases (and character aliases), parse multiple speakers for turns if needed, and others that exist due to the transcripts

[4] dnd.wizards.com/dungeons-and-dragons
[5] criticalrole.fandom.com/wiki/List_of_episodes

Metric	CRD3	MELD	M. WOZ	AMI	CNN	DailyMail
Dialogue Count	159	190	10438	142	92465	219506
Turn Count	398682	13708	143048	79672	3074340	6189038
Total token count in dialogues	5056647	120913	1886018	706803	60476397	154282948
Unique token count in dialogues	42509	6251	20197	9958	341451	596032
Avg. turns per dialogue	2507.4	72.2	13.7	561.1	33.4	28.2
Avg. tokens per turn	12.7	8.82	13.2	8.9	19.7	24.9
Total token count in summaries	327899	-	-	22965	3897045	11308821
Avg. tokens per summary	2062.3	-	-	161.7	42.1	51.5
Avg. summary:dialogue token ratio	0.065	-	-	0.038	0.085	0.087

Table 1: We compare CRD3 with other similar datasets. MELD, Multi-WOZ, and AMI are dialogue datasets. We use the subset of the AMI dialogues with available abstractive summaries. CNN and Daily Mail are abstractive summarization datasets for news articles (we treat an article as a dialogue and a sentence as a turn).

being written over time by fans. We also replace all instances of character aliases in the speaker field with the real speakers' names to reduce noise. Along with the cleaned data, we provide the raw transcription data to document the changes via diff.

3.3 *Critical Role* Episode Summaries

The summaries for each episode were mined from the *Critical Role* Fandom wiki. The summaries are unique in that they are structured and offer different levels of summarization. Most episodes have a (1) wiki opening blurb, which offers briefest level of summarization. This is followed by a synopsis section which is (usually) comprised of several parts: (2) pre-show and announcements, where some logistical information is mentioned; (3) recap, where the previous episode is summarized (usually done by Matt in the episode and is narrative focused); and (4) the episode's plot which is the largest part and summarizes the narrative developments of the episode. The plot sections are also usually divided into sub-sections aligned to narrative topics. Sometimes the wiki also has a break and post-episode sections (usually non-narrative), which we include in the dataset.

3.4 Analysis and Comparison

Refer to Table 1 for turn and token count comparisons. CRD3's total turn count, turns per dialogue, and unique token count are substantially larger than MELD (Poria et al., 2019a) (scripted Friends TV show dataset), Multi-WOZ (Budzianowski et al., 2018) (unscripted task-oriented dialogue dataset), and AMI (Carletta et al., 2006) (unscripted meetings dataset). For AMI, we only consider the dialogues with available abstractive summaries [6]. Multi-WOZ is dyadic while AMI, MELD, and CRD3 have multiple speakers per dialogue.

[6] github.com/gcunhase/AMICorpusXML

We extract 72 total speakers from the entire CRD3 dataset; 9 of which are the main cast (players and DM) and make up 99.48% of the total turns; the DM alone makes up 111,994 turns. In comparison, the 6 main cast of MELD make up 83.27% of the total turns. In addition to real (human) speakers, there are also purely in-game characters role-played by the DM. The indication of the DM role-playing through the use of quotes seem to be mostly consistent in the transcripts. As a loose measure of role-playing, we find the turns that contain quotes from the DM (≈21383) and compare to all other players (≈2497). A core aspect of the game is players querying the DM, so we also measure the instances of questions from a player (turn ending in '?') followed by a DM response; a mean of 199 per dialogue with 58 standard deviation. Finally, we apply the spaCy English NER model on all dialogues as a loose measure of named entity presence. We get a mean of 1275 entities per dialogue with standard deviation of 344.5.

For the summaries, we measure the token counts per summary and compare to AMI, CNN, and Daily Mail (Table 1). Again, CRD3 is substantially larger (though smaller in total tokens than the news datasets). The news datasets also feature more summary-article pairs, making them more amenable to current neural ML approaches; we address this for CRD3 in Section 4. We also measure the compression of the original text to summary via ratio of tokens per summary to tokens per original text and find they correspond to the ratios of total tokens to unique tokens. Finally, we measure the average token count and standard deviation of each section of the structured summaries for the CRD3 dataset (outlined in Section 3.3): (1) Wiki opening blurb: 50 ± 16.7; (2) pre-show and announcements: 183 ± 254; (3) recap: 335 ± 123.9; and (4) episode plot: 1544 ± 1553.7.

4 Scaling up the Dialogue Summaries

The CRD3 dataset can be applied to many tasks, but we find abstractive dialogue summarization the most compelling task to explore in this paper. Due to the extensive length of the dialogues and summaries, and the frequent context switching and noise, we are presented with challenges that are poorly addressed by the current modeling and evaluation methods:

1. The dataset has relatively few episodes (159); as is, this is not enough samples to train, test, and validate using current neural approaches.
2. The current, most successful summarization approaches do not explicitly attempt to capture coreference, semantics, and pragmatics in very long documents or conversations.
3. Current automatic summarization evaluation methods have specific failures in evaluating narrative summarization.

We do not attempt to propose a solution for either the second or third challenges, as they are beyond the scope of this paper. Instead, we address the first challenge by proposing a novel data augmentation method to dramatically scale up the number of available summary-dialogue turn sequence pairs. That outcome enables the community to start modeling and evaluation for the dialogue summarization task and we discuss initial benchmark results over this augmented set in Section 5.

4.1 Data Augmentation via Text Alignment

We found that the summaries written by fans on the wiki are detailed, mostly ordered with respect to the corresponding episode, and mostly non-repetitive. Due to the large number of sentences in the summaries, we can break up the summaries into chunks and align each chunk to some continuous segment of the dialogue. Formally, given dialogue D consisting of T turns $\{t_i | i \in 1 \dots T\}$ and summary S split into n contiguous chunks $\{s_i | i \in 1 \dots n\}$, we try to determine $A = \{a_i | i \in 1 \dots n\}$ where a_i is a contiguous set of turns from D ($a_i = t_{j:k}$) and where t_j and t_k ($j \leq k$) are the earliest and latest turns in D to align to s_i; refer to Figure 2. To determine A, we try two approaches.

Greedy Algorithm We make an independence assumption for all s and t and try to maximize an alignment score, $\alpha(A; S, \beta)$, where $\beta(s, a)$ calculates an alignment score between a single s and a.

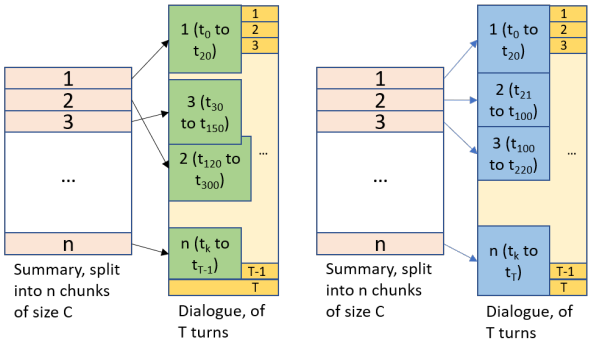

Figure 2: Chunking and mapping of C contiguous summary sentences onto the T turns of the dialogue. The greedy approach (left) has no order or contiguity constraint. The Needleman-Wunsch approach (right) has strict order and contiguity constraints.

$$\alpha(A; S, \beta) = \sum_{i=0}^{n} \max_{\substack{0 \leq c \leq T \\ 0 \leq w \leq 14}} (\beta(s, t_{c-w:c+w})) \quad (1)$$

where bounds for w are determined empirically. For several dialogues, we tested $0 \leq w \leq T$, but this had no change in the final assignments A and greatly increased computation time. To choose β, we tried several scoring functions including variations of ROUGE (Lin, 2004), variations of TF-IDF (Jones, 1988), and other n-gram overlap scorings. We selected a scaled version of ROUGE-F1 score:

$$\begin{aligned} \beta(s, a) &= |\tau(s) \cap \tau(a)| * ROUGE_{F1} \\ &= \frac{2 * |\tau(s) \cap \tau(a)|^2}{|\tau(s)| + |\tau(a)|} \end{aligned} \quad (2)$$

where τ is a tokenization function for the given text. The scaling via $|\tau(s) \cap \tau(a)|$ term gives extra importance to the absolute token overlap count.

To calculate the tokens, we found just unigrams and bigrams gave us the least noisy alignments. We also found lemmatization and stop-word removal greatly reduces the alignment quality because of the large number of n-grams (≥ 2) from the turn windows that are directly used in the summaries.

In Figure 3(a), we plot the turn indices as a function of the summary chunk indices. We notice the greedy alignment approach can largely preserve the order of the summary chunks relative to the dialogue turns, without any ordering constraints. However, there are some issues with this method. First, it allows out-of-order alignments of summary chunks, which we have assessed as almost always erroneous in this dataset. Second, the recall can

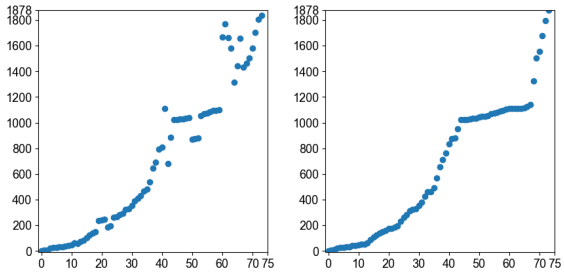

Figure 3: (a) Midpoints of turn sequences as a function of the summary chunk indices for campaign 2 ep. 31, determined by the greedy approach. The plot is generally monotonic, with the out of order points verified as misalignments. After assessing many dialogue and summary pairs, we determined a strong monotonic assumption for this dataset. (b) For the same summary sentence chunk indices as in graph (a), we plot the new turn sequence midpoints as determined by the Needleman-Wunsch approach. The plot is now perfectly monotonic due to the ordering constraint and captures previously missed turn sequences.

be low due to early cutoffs at boundaries, generally because of extensive chit-chat in between two salient utterances. Forcing boundaries between a_i and a_{i+1} to be contiguous leads to lower precision due to salient utterances being incorrectly assigned near the borders of the turn windows.

Needleman-Wunsch Algorithm The recursive approach to determining A involves imposing strict order constraints using the sequence alignment algorithm Needleman-Wunsch (Needleman and Wunsch, 1970), similar to (Nelken and Shieber, 2006). The algorithm imposes order by forcing a_i and a_{i+1} to be assigned to contiguous turn windows. We can also forgo the maximization over some window w as the algorithm does this by virtue of its score maximization function. We tried several functions for β, including the TF-IDF function proposed by (Nelken and Shieber, 2006) and found (2) still performs best. To use the algorithm, we first apply β independently for each turn (of size 1) and summary chunk to generate a match-score matrix M of size $T \times n$. We then build an alignment score matrix H of size $(T+1) \times (n+1)$ using:

$$H_{xy} = max \begin{cases} H_{y-1,x-1} + M_{y-1,x-1} \\ H_{y-1,x} + M_{y-1,x-1} \\ H_{y,x-1} + M_{y-1,x-1} \end{cases} \quad (3)$$

with $M_{y-1,x-1} = \beta(s_{x-1}, t_{y-1})$; $1 \leq y \leq T$; and $1 \leq x \leq n$ and the first column and row of H initialized to $-y$ and $-x$ respectively. We perform the traceback from $H_{T+1,n+1}$ to $H_{0,0}$ to generate

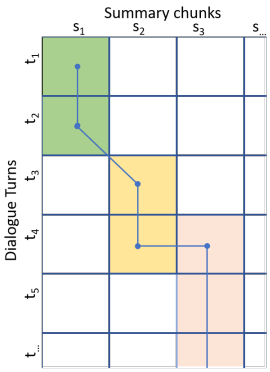

Figure 4: Visualization of the traceback along the H matrix in the Needleman-Wunsch alignment approach. Each vertical line for s_i is the corresponding $a_i = t_{j:k}$.

the alignment A where each $a \in A$ can be seen as a vertical line in the traced path (Figure 4).

We exclude gap penalties when generating H, since we want to allow multiple turns to be assigned to a summary chunk and we want to allow a single turn to overlap several summary chunks. We also notice that column-wise normalization on M reduced the quality of the alignments substantially because large scores can act as an anchor for the algorithm to localize erroneous alignments. It forces the algorithm to 'catch up' or 'pull back' the turn alignments to include the high $M_{y,x}$ in the final path. Normalization also reduces incentives to keep the path going down a column and heavily favors moving to the next column (summary chunk). We can visualize the improvements in Figure 3(b), where we also notice the algorithm captures turns past t_{1833} (upto t_{1878}) that were previously ignored, leading to higher recall – we manually verified this.

The strong ordering constraint is also the source of some noise. For example, if a summary alignment overshoots the correct turn window by a large margin, it is likely that the subsequent summaries will also be misaligned due to the contiguity constraint. However, the localization effect due to large M scores help mitigate this. Another source of noise is the forced alignment of the first and last turns in dialogues that continue past the summary.

We also analyze the distribution of the scores along the paths (each path normalized to 1) traced on M with respect to the nine main players (Table 2). This gives us the distribution of the player contributions to the summaries. Matt's turns contribute most to the summaries since he contributes the most salient narrative points. As the Dungeon Master, he is responsible for world building and the narrative's interaction with the other players. We

Player	β
MATT	0.0307±.0008
ORION	0.0086±.0014
LIAM	0.0083±.0005
TALIESIN	0.0074±.0005
SAM	0.0070±.0004
MARIHSA	0.0058±.0003
TRAVIS	0.0057±.0004
LAURA	0.0056±.0003
ASHLEY	0.0048±.0006

Table 2: Mean (± 0.95 conf. interval) summary contribution scores for each player calculated from the normalized paths traced on M as determined by the algorithm on H.

Chunk Size	w/o Filtering	w/ Filtering
2	18569	11124
3	18438	11635
4	18378	11484

Table 3: number of s_i, a_i pairs generated for each chunk size with and without filtering.

can see the other players have much lower mean scores. One explanation for this is that they engage in more non-narrative chit chat than Matt, which leads to a lower mean β.

Data Augmentation Running the Needleman-Wunsch algorithm for a dialogue D will give us N s, a pairs. We can extend this by calculating S as $S_0 \dots S_{C-1}$ where C is the chunk size and S_x is the shift in the starting point of the contiguous chunking windows. For each of these S_x, we can then determine an A_x pair. This method increases our s, a pairs by a factor of C. We can go further by running this for different chunk sizes. For our experiment, we chose to run this algorithm for C=2, 3, and 4 sentences. We remove dialogues with $|S| \leq 10$ chunks (since there are some incomplete wikis) and get 55385 s, a pairs. To reduce noise, we also: (1) impose $2 < |t_{j:k}| \leq 100$; and (2) strip out pairs where s_i contains "Q: " (signifies a differently formatted question answer segment in an episode). We end up with 34243 pairs (Table 3), a substantial increase from the original 159 summary, dialogue pairs. Refer to Figure 1 and to the Appendix for examples of the summaries and examples. These are then split as 26232 training, 3470 validation, and 4541 testing s, a pairs; refer to Appendix for details.

We calculate precision and recall with respect to the turns on a random sample of 100 pairs from the training split of these 34243 pairs and obtain a precision of 0.8692 and recall of 0.9042. Refer to Appendix for precision and recall calculation

Summary
"The Mighty Nein make their way up the ladder and through the hatch into the Keystone Pub proper, where they order breakfast. A hooded female wearing a long green cloak covering her left face and side approaches and asks if they're heading into the swamp today– she's desperate to go there herself. Calianna apologizes for bothering them, but she couldn't help but overhear their conversation last night."
Factoid Question
1. Who was overhearing the Mighty Nein's conversation the previous night?
Multiple Choice Question
2. What do the Mighty Nein have at the Keystone Pub? **(A)** drinks **(B)** dinner **(C)** lunch **(D)** breakfast

Figure 5: Example of questions constructed for a human-written summary chunk aligned to a set of turns.

method. We find precision errors are mostly from extraneous trailing or leading turns attached to the properly aligned set of turns, and almost never from complete misalignment. We find recall errors are from turn sequences that start too late or end too early, and also almost never from complete misalignment. In most cases where a contains a recall error, we notice the precision for that a is 1.0, because a ends up being a subset of the correct $t_{j:k}$. We posit this is due to the strong order constraints of the algorithm and our post-alignment filtering, which removes the pairs with the highest risk of complete misalignment.

As a measure of quality of the human written summaries, we also perform a question-answering task on a random sample of 50 s_i, a_i pairs from the filtered set. First the questioner records two questions and answers per pair, with the questions and answers coming only from the summaries s_i. For each pair, there is one factoid question with an open-ended answer and one multiple choice question with four possible answers. The factoid question can be answered by yes—no responses, entity names, or short text. The multiple choice question has at most one correct answer of the four contained in the summary chunks. (Figure 5). The questions are then answered by another person, using only the aligned turns a_i from the pair.

The scores are recorded in Table 4. Out of the 19 incorrect answers, we found that 17 of them were due to summary alignment errors. This is where the correct information was in the dialogue, but not in the aligned set of turns. The other 2 were due to misinterpretation of the question when answering. This indicates, with perfect alignment, all questions

Question Type	Correct	Incorrect	Precision
Free Form	39	11	78%
Multiple Choice	42	8	84%
Total	81	19	81%

Table 4: Correct and incorrect answers for the Q&A evaluation method, for measuring precision w.r.t. the human written summaries in the s_i, a_i pairs.

could have been answered correctly; meaning what is in the summaries is an accurate reflection of what is in the transcript. However, we recognize all the information in the transcripts is not necessarily in the summaries; for example, out-of-game information. We also notice that multiple choice questions have a higher accuracy due to easier questions and additional context provided by the set of answers themselves, and not due to random guessing. We also found that 12 incorrect answers were due to no answer, meaning the answerer did not feel they had enough information to attempt an answer. For the other 7, the answerer felt that at least some information pertaining to the question was available in the aligned turns.

Unlike ROUGE precision, which relies on word overlap, this evaluation can incorporate latent semantic and contextual information. It is important to note that latent information used when answering varies greatly between people, making this method subjective with respect to the answerer. In future work, it would be interesting to measure variance of accuracy and information in the answers using a large number of people.

5 Summarization Benchmark Results

5.1 Benchmarking Approach

We establish a baseline for abstractive summarization by using the neural summarization architecture introduced by (Chen and Bansal, 2018)[7]. The generated data has noise due to imperfections in the alignment method and due to potentially broken coreference, so we use the model in a semi-supervised fashion.

We choose this architecture as a baseline for several reasons: (1) The paradigm for narrative summarization from noisy dialogue is close to the paradigm assumed by Chen and Bansal. Namely, first extract salient sentences, then abstractively rewrite them with an included copy mechanism to deal with OOV words. (2) The ability to analyze the extractor behavior separately from the abstractor due to the independence of training (before connection by the reinforcement learning mechanism). (3) The speed of training due to the shortened input-target pairs.

[7] github.com/ChenRocks/fast_abs_rl

	R1	R2	RL	M
Extractive (rnn-ext + RL)				
P	20.83±.34	**7.34**±0.28	18.38±.32	
R	**44.59**±.66	**17.42**±.62	**39.22**±.61	**16.58**
F1	**25.20**±.34	**9.23**±.32	**22.20**±.32	
Reported Metrics on CNN/DM				
F1	41.47	18.72	37.76	22.35
Abstractive (rnn-ext + abs + RL + rerank)				
P	**27.38**±.34	5.91±.20	**25.18**±.32	
R	22.65±.27	4.75±.16	20.74±.26	8.33
F1	23.35±.23	4.91±.16	21.41±.23	
Reported Metrics on CNN/DM				
F1	40.88	17.80	38.54	20.38

Table 5: ROUGE (Precision, Recall, F1 ± 0.95 conf. interval) and METEOR (M) metrics on the CRD3 test set using the purely extractive and extractive+abstractive architecture proposed by Chen and Bansal. We show the metrics on the CNN/Daily Mail dataset for the same models as reported by Chen and Bansal.

We briefly describe the model: First, the model optimizes a sentence extraction module and an abstractive rewrite module independently using maximum-likelihood objectives. Then, end-to-end training is achieved by applying policy gradient methods (due to the "non-differentiable hard extraction" performed by the extractor). The extractor uses a temporal convolutional model to obtain hierarchical sentence representations, then selects sentences using a pointer network. The abstractor is an encoder-aligner-decoder network with a copy mechanism for OOV words. Due to the large amount of non-narrative chit-chat turns between salient turns, we train the extractor on a sequence of turns rather than individual sentences.

5.2 Evaluation and Analysis

We use precision, recall, and F-1 scores of ROUGE-1, 2, and L, along with METEOR (Denkowski and Lavie, 2014) to evaluate the generated summaries (Table 5). We run these metrics on the test set, using both the combined extractive-abstractive model and the purely extractive model for analysis on what turns are considered salient.

The purely extractive model significantly outperforms the combined model in recall and in F-1, due to the much higher recall. In the validation set, we notice the recall measures are improved by the n-grams in summary chunks that have indirect speech ("fjord says", "he says", etc). In the validation

Generated Abstractive Summary
he says he feels worried about trying to learn about these abilities and abilities .
he asks if she could try and cheer .
the group then heads to the tavern .
she asks if she can see jester , and she says she 's really disturbing .

Figure 6: Extractor+Abstractor output for the dialogue sample in Figure 1

set, the mean ratio of unique overlapping summary n-grams to total unique summary n-grams are: 1-gram= 0.679, 2-gram= 0.336, and 3-gram= 0.205. This high rate of 3-gram overlap motivates changes to the modeling architecture that are more lenient towards phrasal copy instead of just enabling word copy and depending on the learned language model and the word level copy probability.

The grammatical person shift and significant paraphrasing of turns lower the precision of the purely extractive model, leading to a higher precision in the combined model. For example in Figure 1, "beau asks about jester ." from the human-authored summary is entirely from turn 1, but the only overlapping word is "jester". From Figure 6, we can see the encoder-decoder model learns the grammatical shift behavior but doesn't include the proper nouns, so the resulting summary misses important speaker information that is included in the human generated summaries. For example, Beau is the character alias for Marisha, which is latent information that was not available to the model at the time of decoding/generation. We also note the encoder-decoder module's learned language model is biased by the narrative elements present in the training dialogue chunks. This causes decoding of similar, but fundamentally different, narrative focused turns to be noisy and nonfactual.

Compared to news summarization metrics with the same model architectures, the dialogue summarization metrics are substantially lower. The disparity in model performance can be attributed to content selection differences between news – where effective summary information is available early in an article (position bias) – and dialogue – where the positional effects are not observed. Other factors include the grammatical and stylistic differences explored earlier. Our findings also confirm the findings of (Kedzie et al., 2018), which compares content selection methods for summarization across various domains (CNN/DM, NYT, DUC, Reddit, AMI, and PubMed). They find a similar disparity in R-2 (recall) and METEOR scores between the news domain and the AMI meeting dialogue domain. They also include an oracle measurement as a performance ceiling; it achieves a max METEOR score of 17.8 and R-2 recall of 8.7 on the AMI corpus. Though ROUGE and METEOR are more useful for relative measurements than absolute, we find the current evaluation methods in summarization lead to skewed and less informative scores in dialogue domains. The problem is compounded in narrative summarization due to narrative specific lexical information, including speaker aliases. For example, METEOR specifically considers synonyms, paraphrases, and function words; all of which can change a lot from narrative to narrative.

6 Conclusion and Future Work

Dialogue understanding and abstractive summarization remain both important and challenging problems for computational linguistics. In this paper, we contribute the *Critical Role* Dungeons and Dragons Dataset (CRD3), a linguistically rich dataset with dialogue extracted from the unscripted, live-streamed show *Critical Role* and long, abstractive summaries extracted from the *Critical Role* Fandom wiki. We provide a data augmentation method to help the community start modeling and evaluation for the dialogue summarization task and discuss the initial modeling benchmark results. We find current paradigms in summarization modeling to have specific failures in capturing semantics and pragmatics, content selection, rewriting, and evaluation in the domain of long, story-telling dialogue. We hope CRD3 offers useful, unique data for the community to further explore dialogue modeling and summarization. We also hope that the dataset can be added to in the future with multi-modal extractions, more granular annotations, and deeper mining of the wiki.

Acknowledgments

First and foremost, we thank the Critical Role team[8] for creating a fun, entertaining, organized, and growing set of livestreams that we used in this dataset. Next, we thank the CRTranscript team[9] for providing high quality transcripts of the show for the community and we thank all the contributors of the Critical Role Wiki. Finally, we thank Rahul Jha for providing feedback and Oli Bailey for contributing evaluation questions.

[8] critrole.com/team

[9] crtranscript.tumblr.com/about

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, et al. 2012. Developing a corpus of strategic conversation in the settlers of catan.

Rafael E. Banchs. 2012. Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–207, Jeju Island, Korea. Association for Computational Linguistics.

Alan W Black, Susanne Burger, Alistair Conkie, Helen Hastie, Simon Keizer, Oliver Lemon, Nicolas Merigaud, Gabriel Parent, Gabriel Schubiner, Blaise Thomson, et al. 2011. Spoken dialog challenge 2010: Comparison of live and control test results. In *Proceedings of the SIGDIAL 2011 Conference*, pages 2–7. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39, Berlin, Heidelberg. Springer-Verlag.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

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

Alex Djalali, Sven Lauer, and Christopher Potts. 2012. Corpus evidence for preference-driven interpretation. In *Proceedings of the 18th Amsterdam Colloquim Conference on Logic, Language and Meaning*, AC'11, pages 150–159, Berlin, Heidelberg. Springer-Verlag.

Pierfranca Forchini. 2009. Spontaneity reloaded: American face-to-face and movie conversation compared. In *Proceedings of the Corpus Linguistics Conference 2009 (CL2009),*, page 400.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 968–974, Brussels, Belgium. Association for Computational Linguistics.

John J. Godfrey, Edward Holliman, and Jan McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520 vol.1.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Denver, Colorado. Association for Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2018. What's this movie about? a joint neural network architecture for movie content analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1770–1781, New Orleans, Louisiana. Association for Computational Linguistics.

Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *IVA*.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Zhichao Hu, Michelle Dick, Chung-Ning Chang, Kevin Bowden, Michael Neff, Jean Fox Tree, and Marilyn Walker. 2016. A corpus of gesture-annotated dialogues for monologue-to-dialogue generation from personal narratives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3447–3454, Portorož, Slovenia. European Language Resources Association (ELRA).

Hayley Hung and Gokul Chittaranjan. 2009. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In *ACM Multimedia*.

Karen Spärck Jones. 1988. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 60:493–502.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Geoffrey Leech. 1992. 100 million words of english: the british national corpus. *Language Research*, 28(1):1–13.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713, New Orleans, Louisiana. Association for Computational Linguistics.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn Walker. 2015. Using summarization to discover argument facets in online idealogical dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 430–440, Denver, Colorado. Association for Computational Linguistics.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Dont give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Haojie Pan, Junpei Zhou, Zhou Zhao, Yan Liu, Deng Cai, and Min Yang. 2018. Dial2desc: End-to-end dialogue description generation. *arXiv preprint arXiv:1811.00185*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, Lisbon, Portugal. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer.

Preethi Vaidyanathan, Emily T. Prud'hommeaux, Jeff B. Pelz, and Cecilia O. Alm. 2018. SNAG: Spoken narratives and gaze dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 132–137, Melbourne, Australia. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *Proceedings of the International Conference on Machine Learning, Deep Learning Workshop*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

A Appendices

A.1 Summary Dialogue Alignment Precision and Recall Calculation Method

We calculate precision and recall for summary dialogue alignment with respect to the dialogue's turns in Section 4.1. Here, we describe our method for calculating precision and recall.

Precision is expressed as a function of true positives and false positives and recall is expressed as a function of true positives and false negatives. For each alignment $a_i \in A$, we classify each of its turns t as a True Positive (TP), False Positive (FP), or False Negative (FN). We take the counts of all TP, FP, and FN over the entire A and perform the precision and recall calculations, precision=$\frac{total(TP)}{total(TP)+total(FP)}$, recall=$\frac{total(TP)}{total(TP)+total(FN)}$.

A.1.1 TP, FP, FN Classifications

We have the following guidelines to classify a turn in a_i as a TP, FP, or FN.

1. First, find the earliest and latest turns in the original dialogue that correspond to the summary chunk s_i. All alignments $a \in A$ are a contiguous sequence of turns extracted from the dialogue. For example, in the summary chunk in Figure 1, the earliest turn in the entire dialogue that corresponds to the summary is (1) in the alignment. The latest turn in the entire dialogue that corresponds to the summary is (7) in the alignment (we verify this by looking at the turns in the original dialogue before and after the sequence presented in the alignment).
2. Any turn in the alignment in between the earliest and latest turns identified in Step 1 (inclusive) is considered a true positive. Any turn in the alignment outside of the earliest and latest turns identified in Step 1 is considered a false positive. In Figure 1, turn (0) would be considered a false positive because it does not correspond to any of the summary sentences (0,1,2,3). Turns (1,2,3,4,5,6,7) are considered true positives since they are between the earliest and latest turns that correspond to the summary sentences in original dialogue.
3. Any turn between the earliest and latest turns identified in Step 1 that is NOT present in the alignment is considered a false negative. In Figure 1, if the turn (7) was not in the alignment, it would be considered a false negative because the turn (7) corresponds to the summary sentence (2) and is between the earliest and latest turns identified in Step 1 (turns 1 and 7 respectively).

A.2 More Examples of Summary-Dialogue Alignments

We give more examples of summary-dialogue alignments (s_i, a_i) pairs. For the sake of brevity, we chose to show examples that were only 10 turns or smaller. Please refer to the dataset itself for much longer samples.

In Figure 7, we have an alignment with a large recall error. In Figure 8, we have an example of a summary referring to out-of-game turns. We find these types of summaries are typically written for break-times in the show, before the start of a game session, or after the end of a game session. Generally, they seem to make up a smaller portion of the overal summary content. This example in particular is for a Q/A session the team held after their session[10]. In Figure 9, we have a perfect alignment, with the summary explicitly capturing implied information in the turns. There are also examples

[10] Attack_on_the_Duergar_Warcamp episode

	Recall Error Dialogue Chunk
0	MATT: "End of your turn, it's going to use two actions to do a wing attack, beating its wings, hitting every creature within 15 feet. You're out of range, actually, Marisha. Grog, I need you to make a dexterity saving throw."
1	TRAVIS: "I think I have advantage on this because of rage. I do. 21."
2	MATT: "21? That unfortunately fails. You take 15 points of bludgeoning damage, and you're knocked prone. Also, Pike and Vax, you both fail a death saving throw from the bludgeoning winds of the ice dragon's wings beating downward."
	Aligned Summary Chunk
0	"Scanlan takes a Greater Healing Potion and moves towards Vorugal. He hits him with a Fireball."
1	"Vorugal uses a wing attack against Grog, hitting both Vax and Pike as well, losing a death save each."

Figure 7: A (not tokenized) turn sequence and the associated human written summary chunk after the text alignment process. It is clear from the second sentence of the summary chunk, that the turn aligned turns are a subset of the the true turn sequence the summary chunk is referring to. In order to capture the turns referred to by the first sentence in the summary, we need to include the additional 29 preceding turns in the dialogue (which are treated as 29 False Negatives).

	Out of Game Dialogue Chunk
0	ORION: "Ooh, like Thai food."
1	LIAM: "I like Indian."
2	MATT: "Ooh, Indian is good."
3	ASHLEY: "I really noticed–"
4	ZAC: "Let them know not to order food."
5	LIAM: "Don't, that's a terrible idea."
6	ORION: "We just had a bunch of chicken."
7	MARIHSA: "Oh you mean like right now? Yeah, don't do it right now."
8	ZAC: "If you tell them what you want, all of a sudden I'll get a call, like, "your food is on the way!""
	Aligned Summary Chunk
0	"Liam, Matt, Marisha, and Taliesin like Indian food."
1	"Zac chimes in telling the chat not to order any more food right now."

Figure 8: An out-of-game turn sequence and summary chunk. We find a single precision error in this alignment with Orion mentioning Thai food, which is not in this summary chunk.

of role-playing by Matt in this turn sequence, as he speaks to the other players from the perspective of the in-game character Ripley. This is shown through the use of quotes in turns 0, 4, and 6.

A.3 Train, Validation, Test Split Method

In Section 4.1, we split the aligned 34243 pairs into 26232 training, 3470 validation, and 4541 testing pairs. Here, we briefly describe our method.

We first split the 159 dialogues into an (80%, 10%, 10%) train, validation, and test split based on order. This guarantees that episodes from validation will succeed episodes in training, and episodes in testing will succeed episodes in validation. We take all the s, a pairs from these dialogues and put them into their respective train, validation, test sets. We chose to split by this method so that (1) there will never be an episode that is in more than one train/val/test set; (2) no summary of chunk size C_i from validation or testing is a subset of summary of chunk size C_j from the training set where $i \leq j$, thus avoiding bias in the final metrics; and (3) we can train on information that happened in the show prior to information we validate or test on, thus better mimicking a real-world scenario where you cannot train on future information.

	In Game Dialogue Chunk with Roleplay
0	MATT: " "I don't spend my time wondering or curious about her well-being! I just know that she is usually here." "
1	TALIESIN: "Anna. I'm going to take a leap of faith and believe, contrary to all evidence, that you are a smart woman. I pull out the gun, and I put it to her head. Now. If you were the Briarwoods, where would you put my sister?"
2	LAURA: "An important question here, Percy. Are they keeping her, or is she here of her own volition?"
3	TALIESIN: "I don't know. And if you don't know, make me believe it."
4	MATT: " "I know she's not allowed anywhere near the ziggurat or near our distillery." "
5	TALIESIN: "Distillery? I pull the gun away."
6	MATT: "She breathes a sigh of relief. "That's been largely my project as part of this entire endeavor. All right, so when I was brought in here, I was tasked to experiment with the design and create large amounts of a very, very delicately prepared acidic compound, one that could dissolve the stone of your whitestone and distill it down into pure residuum. This would allow the bulk creation of a very powerful magical essence for use in construction materials that we could instill and use apparently for this ziggurat, as well as other such things. Thus, that was my main reason for being here. We were ahead of schedule, and I completed the bulk of our development weeks ago, and I no longer had much of a purpose here." "
	Aligned Summary Chunk
0	"When asked where she could be, Ripley claims that she prefers not to pay attention to the well-being of others, only that she is usually in her room. Percy then starts to lose his patience."
1	"Giving in to Percy's threat, Ripley mentions that Cassandra is not allowed anywhere near the Ziggurat or the "distillery". "
2	"He lowers the weapon to allow her to explain."

Figure 9: An turn sequence and summary chunk with perfect alignment. We observe there is implied information in the turns that is captured more explicitly in the summaries. For example "Giving into Percy's threat, Ripley..." summarizes what happens after turn 1 where Ripley is threatened with the gun and "gives in" by answering Laura's question.

As new Critical Role episodes and seasons are added, we hope to expand the CRD3 dataset correspondingly. Future work might include splitting the training, validation, and testing sets based on season or some method that guarantees independence between narrative elements from the summaries and turns in the training, validation, and testing sets. Note, as new Critical Role episodes are added, we will keep the original version preserved so as to keep the experiments and analysis reproducible.