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
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\text{critrole.com}
2\text{criticalrole.fandom.com}
3\text{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 under-explored 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), (Lis et al., 2015), (Louis and Sutton, 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 storytelling. 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 wiki5. 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

---

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>CRD3</th>
<th>MELD</th>
<th>M. WOZ</th>
<th>AMI</th>
<th>CNN DailyMail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue Count</td>
<td>159</td>
<td>190</td>
<td>10438</td>
<td>142</td>
<td>92465</td>
</tr>
<tr>
<td>Turn Count</td>
<td>398682</td>
<td>13708</td>
<td>143048</td>
<td>79672</td>
<td>3074340</td>
</tr>
<tr>
<td>Total token count in dialogues</td>
<td>5056647</td>
<td>120913</td>
<td>1886018</td>
<td>706803</td>
<td>60476397</td>
</tr>
<tr>
<td>Unique token count in dialogues</td>
<td>42509</td>
<td>6251</td>
<td>20197</td>
<td>9958</td>
<td>341451</td>
</tr>
<tr>
<td>Avg. turns per dialogue</td>
<td>2507.4</td>
<td>72.2</td>
<td>13.7</td>
<td>561.1</td>
<td>33.4</td>
</tr>
<tr>
<td>Avg. tokens per turn</td>
<td>12.7</td>
<td>8.82</td>
<td>13.2</td>
<td>8.9</td>
<td>19.7</td>
</tr>
<tr>
<td>Total token count in summaries</td>
<td>327899</td>
<td>-</td>
<td>-</td>
<td>22965</td>
<td>3897045</td>
</tr>
<tr>
<td>Avg. tokens per summary</td>
<td>2062.3</td>
<td>-</td>
<td>-</td>
<td>161.7</td>
<td>42.1</td>
</tr>
<tr>
<td>Avg. summary:dialogue token ratio</td>
<td>0.065</td>
<td>-</td>
<td>-</td>
<td>0.038</td>
<td>0.085</td>
</tr>
</tbody>
</table>

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. Multi-WOZ is dyadic while AMI, MELD, and CRD3 have multiple speakers per dialogue. 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.

We 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.
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 $\{t_i | i \in 1 \ldots T\}$ and summary $S$ split into $n$ contiguous chunks $\{s_i | i \in 1 \ldots n\}$, we try to determine $A = \{a_i | i \in 1 \ldots n\}$ 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 $a$.

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

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 $\beta$, 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:

$$\beta (s, a) = |\tau (s) \cap \tau (a)| \ast ROUGE_{F1}$$

$$= \frac{2 \ast |\tau (s) \cap \tau (a)|^2}{|\tau (s)| + |\tau (a)|}$$

where $\tau$ 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 ($\geq 2$) 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
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 contiguous 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 $\beta$, including the TF-IDF function proposed by (Nelken and Shieber, 2006) and found (2) still performs best. To use the algorithm, we first apply $\beta$ 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:

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

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 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
We end up with 34243 pairs (Table 3), a substantial Wunsch algorithm for a dialogueokus) and get 55385 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 $H$.

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

<table>
<thead>
<tr>
<th>Chunk Size</th>
<th>w/o Filtering</th>
<th>w/ Filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>18569</td>
<td>11124</td>
</tr>
<tr>
<td>3</td>
<td>18438</td>
<td>11635</td>
</tr>
<tr>
<td>4</td>
<td>18378</td>
<td>11484</td>
</tr>
</tbody>
</table>

We 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 $\beta$.

**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 $S_0 \ldots S_{C-1}$ 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 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

<table>
<thead>
<tr>
<th>Player</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATT</td>
<td>0.0307±0.0008</td>
</tr>
<tr>
<td>ORION</td>
<td>0.0086±0.0014</td>
</tr>
<tr>
<td>LIAM</td>
<td>0.0083±0.0005</td>
</tr>
<tr>
<td>TALIESIN</td>
<td>0.0074±0.0005</td>
</tr>
<tr>
<td>SAM</td>
<td>0.0070±0.0004</td>
</tr>
<tr>
<td>MARIHSA</td>
<td>0.0058±0.0003</td>
</tr>
<tr>
<td>TRAVIS</td>
<td>0.0057±0.0004</td>
</tr>
<tr>
<td>LAURA</td>
<td>0.0056±0.0003</td>
</tr>
<tr>
<td>ASHLEY</td>
<td>0.0048±0.0006</td>
</tr>
</tbody>
</table>

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 $H$.

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

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

**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

“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.”

**Summary**

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...
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 Summary 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 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.

\(^7\)github.com/ChenRocks/fast_abs_rl

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free Form</td>
<td>39</td>
<td>11</td>
<td>78%</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>42</td>
<td>8</td>
<td>84%</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>19</td>
<td>81%</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>RI</th>
<th>R2</th>
<th>RL</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extractive (rnn-ext + RL)</td>
<td>(P) 20.83±.34</td>
<td>(R) 17.42±.62</td>
<td>39.22±.61</td>
<td>16.58</td>
</tr>
<tr>
<td></td>
<td>(P) 20.83±.34</td>
<td>(R) 17.42±.62</td>
<td>39.22±.61</td>
<td>16.58</td>
</tr>
<tr>
<td></td>
<td>(P) 22.65±.27</td>
<td>(R) 27.38±.34</td>
<td>44.59±.66</td>
<td>25.18±.32</td>
</tr>
<tr>
<td></td>
<td>(P) 22.65±.27</td>
<td>(R) 27.38±.34</td>
<td>44.59±.66</td>
<td>25.18±.32</td>
</tr>
<tr>
<td></td>
<td>(P) 23.35±.23</td>
<td>(R) 27.38±.34</td>
<td>44.59±.66</td>
<td>25.18±.32</td>
</tr>
<tr>
<td></td>
<td>(P) 23.35±.23</td>
<td>(R) 27.38±.34</td>
<td>44.59±.66</td>
<td>25.18±.32</td>
</tr>
</tbody>
</table>

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.

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
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


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


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,

\[
\text{precision} = \frac{\text{total}(TP)}{\text{total}(TP) + \text{total}(FP)}, \quad \text{recall} = \frac{\text{total}(TP)}{\text{total}(TP) + \text{total}(FN)}.
\]

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 that correspond to the summary sentences in original dialogue.

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 overall 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
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 from validation or testing is a subset of summary in training set $C_i$; (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.
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.