Extractive Summarization Techniques with the Critical Role Dungeons and Dragons Dataset (CRD3)

Emma Waddell, Jordan Grayson, Arjun Jaishankar
Department of Computer Science
New York University
{erw364, jtg400, asj397}@nyu.edu

Abstract

In this paper, we describe our applications of extractive summarization techniques towards the Critical Role Dungeons and Dragons Dataset (CRD3). Critical Role is an unscripted, live-streamed show where a group of individuals play Dungeons and Dragons 5th Edition, an open-ended tabletop role-playing game. Using this dataset, collected from 398,682 turns, we summarize episodes or shorter scenes into concise summaries using extractive models, where they are evaluated based on ROUGE-1 and ROUGE-L scores. In addition, we utilize an Ad-Hoc Information Retrieval system using TF-IDF weights and cosine similarity score to retrieve information that assists Dungeons and Dragons 5th Edition world-building and narrative creation, based on a defined series of user queries.

1 Introduction

Automatic text summarization is focused on generating a short and concise summary for a given set of information. Human summarization is often accurate, readable and succinct, and carefully preserves the information and context of the source information. To have an automatic text summarization approach this paradigm is a very challenging task, with current solutions still being far away from existing human performances (Allahyari et al., 2017). However, one of the existing models of machine summarization, extractive summarization, is still quite effective in this field. Extractive summarization methods select sentences directly from the original information source based on its contextual importance. Using these chosen sentences, an aggregate summary is generated (Xiao et al., 2019). Therefore, extractive techniques are often relatively fluent and grammatically correct, although sometimes lacking in coherence and brevity.

This model is particularly useful when it comes to summarizing information that contains fictitious terms and vocabulary. In this paper, we will discuss our exploration of applying the extractive summarization model to a corpus of fantasy-related data, known as the Critical Role Dungeons and Dragons Dataset (CRD3). Recent works on extractive summarization have been rather successful in generating summaries of short news documents (around 650 words/document) (Nallapati et al., 2016) by applying neural Seq2Seq models (Cheng et al., 2016). However when it comes to unique terms, such as fictional city names, monsters and magic, it can be difficult to correctly summarize and weigh information, especially while detaching the summarization from factual information that would otherwise be weighed and summarized with greater priority and precision. This problem is especially prevalent when information is presented in dialogue, and not as prose. Lengthy dialogues about fictional events cover a variety of topics, with a variety of inflections, perspectives and interpretations. Accurately and concisely summarizing these factors is a core problem we intend to address over the course of this paper.

1.1 *Critical Role* and CRD3

The Critical Role show is a weekly unscripted live stream of a fixed set of individuals playing Dungeons and Dragons 5th Edition, a popular table-top role-playing game (Rameshkumar et al 2020). The show is set in a fictional world, filled with creatures, histories and characters inspired from mythology and popular culture. During an episode, the eight players and their Dungeon Master (DM), Matthew Mercer, attempt to fulfill multiple objectives in the game's fictional world, explicitly detailing what their characters would do in this imaginary setting.

The Critical Role Dungeons and Dragons Dataset is a set of compiled dialogue transcriptions of each episode, supplemented with online knowledge from the show's fan-base. This dataset contains the following: (i) human-authored summaries of each of the 146 episodes of the show, (ii) extractive summaries of segmented portions of each episode, (iii) complete dialogue transcript of each character within the episode. This dataset will serve as a corpus used to develop, train and test our models and results.

1.2 Contributions

The core contributions of this project are as follows: (i) in order to summarize information that pertains to uncommon topics and/or dialogue based material, we are to create an effective extractive model to be trained on the CRD3. (ii) We intend to test our models on ROUGE-1 and ROUGE-L evaluation schemes and assess whether the results support our goal of effective document summarization. (iii) Utilize existing information weight techniques incorporated in our extractive summarization models to supplement user information retrieval for further information creation in similar contexts.

2 Extractive Summarization

Summarization systems take in a long text document, and output a shortened summary. Extractive systems generate a summary made up of sentences already included in the document. One of the most common and simplest implementations of extractive summarization is a sentence ranking system. The document is scanned, and the frequencies of each word in the document is stored. The sentences with the most high frequency words receive the highest ranking, and the highest ranked 3-5 sentences are then outputted. This system is very simplistic, however, so many systems use TF-IDF (term frequency inverse document frequency) instead of just term frequency, along with other modifications.

3 Related Works

As an improvement for especially long text documents, Wen Xiao and Giuseppe Carenini had their summarizer incorporate the context of the entire document, as well as the local context of the current topic. This is interesting when taken with the CRD3, since within each episode (or entire document) the dialogue is split into turns (or the current topic). This connection represents a possible further improvement to the current system.

Another relevant improvement is Rada Mihalcea's system *TextRank* which is not language specific, and therefore is good within "new languages or domains." Since the CRD3 has so much unique vocabulary, specific to fantasy worlds and tabletop games, *TextRank* helped provide a further enhancement to our system. It described a system in which every sentence is a node on a graph, and the sentences with the most connections to other sentences are ranked higher due to being more central. These ideas led to the implementation of cosine similarity in our system to find the most central sentences in a similar way.

4 Data Analysis and Processing

4.1 Dungeons and Dragons

Dungeons and Dragons is a popular tabletop role-playing game by Gary Gygax and Dave Arneson. With foundations in structured storytelling, players in the game create characters as proxies to interact with the fictional world created by their Dungeon Master (DM). By using their actions, in the form of dialogue and dice rolls, players can explore the environment, talk with fictional characters, battle monsters, and much more (Rameshkumar et al 2020).

4.2 The Structure of the CRD3

The CRD3 consists of 159 episodes worth of dialogue, formatted in a variety of frameworks based on user requirements. Based on our outlined requirements, we chose to use the processed cleaned-data files. Each file consists of the following: (i) episode synopsis, (ii) segment synopsis, (iii) all dialogue linked to the associated cast member. Each file can be up to 30000 lines long.

4.3 Data Preprocessing

To support greater accessibility of data, as well as performance, we chose to parse and reorganize CRD3 information into processed javascript object notation files that can be found within the project's framework. Each file contains a series of hash tables that allow for the organized retrieval of an episode's synopses, dialogue-phrase by index, linked cast member by index, or a formatted construction of the entire episode's dialogue.

5 Difficulties with the CRD3

The CRD3 corpus is complicated to summarize, for three main reasons. For one, a large portion of the text in the corpus represents dialogue that is not part of the game play (greetings, cast introductions, sponsorships, giveaways, etc). The summary found from the show wikipedia, however, only summarizes what happened in the game. Therefore the extractive summarizer should only extract sentences from the in game dialogue, so in the data used by the system the "out of game" turns should be removed.

Also, similarly, terms that only occur in game, and never out of game, are more important and should be weighted higher. This is because those terms are fantasy game specific, so having those terms have a higher weight will help to ensure that no "out of game" sentences are included in the summary.

Finally, Critical Role is quite long, as each episode is four to five hours. A document that long will always be more difficult, but an added difficulty is that since it is a story game, a lot of plot and story progression happens during that time which is complicated to condense into a 4-5 sentence summary. The output is often disjointed, with sentences that are not very well related to each other.

6 Improvements and Methodology

To make an extractive summarizer that worked with this unique corpus, a few improvements were implemented. The original system was a sentence ranking system that used TF-IDF to rank the top four sentences, as previously mentioned.

The first improvement was implemented in the program that parsed and cleaned the .json files. First, an array of “bad” terms was created, based on terms that would only be said while out of game. These included the actors’ names, their Youtube channel (Geek and Sundry), the name of the show (Critical Role), and the name of the game (Dungeons & Dragons). All the turns that included any of these terms were put into a separate text document. Then, the top 10 terms (that were not stop words) in that document were also added to the array of bad terms. That way, most of the turns that occurred out of the game would not be included in the episode text, and therefore would not be included in the extractive summary.

Once this document was created, it was also used in another improvement to the system. Words that were only mentioned in one of the in game turns, and were never mentioned in the out of game turns, were put into a “good” term array. Then, while calculating the weights of each word in the corpus, “good” terms were rated higher than those that were spoken out of game.

The final improvement was implemented cosine similarity to rank the sentences instead of just TF-IDF. That way, the most central sentences within each episode, or the sentences that are similar to the highest number of other sentences in the episode, will be ranked higher. This is helpful to increase cohesion within the extractive summary, since the words in the sentences will be related to each other. A future possible extension to this could be to rank the similarities of sentences within turns as higher than the similarities of sentences between two different turns. This takes ideas from another system that uses both local and global contexts, described earlier in Related Works.

7 Evaluation

Based on the methodology of Revanth and Bailey, as well as other systems related to summarization, we are using a ROUGE evaluation system to analyze our results (Revanth, & Bailey, 2020). Fortunately, every episode of Critical Role has been summarized by fans for the show’s wikipedia page. This gives us a great metric to compare our system to. The ROUGE system stands for Recall-Oriented Understudy of Gisting Evaluation. ROUGE is described as a set of metrics for evaluating automatic summarization of texts as well as machine translation (Lin, 2004). This focus on summarization makes ROUGE a great fit for evaluating our own system.

We have calculated Precision, Recall, and FMeasure scores at various points of development for our system to track our progress throughout development. In the context of our ROUGE evaluation system, the Recall score demonstrates what portion of the reference summary our system is capturing, the Precision score reveals how much of the system summary was relevant or needed, and the FMeasure score reconciles these values together to give a more holistic depiction of the performance of our system.

7.1 ROUGE Implementation

There are various ROUGE metrics available for implementation, and we have employed both a ROUGE-1 and ROUGE-L system. The ROUGE-1 system compares the overlap of unigrams between our summary and the reference summary, and the ROUGE-L system determines the longest matching sequence of words using LCS or Longest Common Substring. The reasoning behind implementing these two specific evaluation systems is because ROUGE-1 focuses on individual unigrams, it offers a finer granularity to its evaluation. Lin describes ROUGE-1 as able to “also show the fluency of the summaries or translation” because if the system summary closely follows the words of the reference summary, it can be a sign that it is more fluent (Lin, 2004). We elected to implement our ROUGE-L in conjunction with our ROUGE-1 system as it helps us determine the quality of our summary on a sentence level. As ROUGE-L also already includes the longest in-sequence common n-grams between our system and the reference summary, it also serves to take the place of a potential ROUGE-N system - which would compare higher order n-gram overlap.

7.2 Abstractive Comparison

A note about summarization systems in general with respect to evaluation is that we found that summarization systems, like translation systems, do not score highly in any evaluation system when looking at the results objectively. For example, our highest scoring ROUGE-1 FMeasure result for an individual episode is .0992. When considering that FMeasure is determined on a zero to one scale, this result may look quite low, however, when evaluating these results with the range that summarization systems generally score, this score is actually relatively good. In order to provide a comparable reference point to our scores, we also ran the CRD3 through a pre-trained abstractive summarization system, the T-5. This was also due to the fact that we originally were interested in pursuing abstractive summarization as a goal of this paper, and see it as a sort of next step to our extractive system. The results of the T-5 system are helpful though because again, evaluated objectively on a zero to one scale, they are quite low. Because of the advanced technology and the longer development time period and fewer budget constraints that this professional system has, it performed better than our own, but this allows us to look at the T-5 results as a sort of upper-bound to our scale. We can think of our ROUGE FMeasure results not only on a zero to one scale, but also -more interestingly- on a zero to T-5 score scale.

7.3 Baseline

In addition to comparing our system and the T-5 pre-trained abstractive system to the reference summary, we developed a baseline system as another comparison point. This naive system involves parsing the corpus chunks, and comparing each word to a POS dictionary to find one noun and one verb for each sentence. This results in baseline summaries that resemble the following excerpt:

> “life is we had part closure had part chest is everyone episode voice moment way he he knew we get announcements campaign…”

This is useful for a variety of reasons. First, being that this is a naive extraction of words, it scores relatively well in terms of ROUGE-1 Recall. Many of the unigrams it draws from the text are often heads of noun and verb phrases and end up in the reference summary. This gives us a meaningful baseline goal to try and surpass with this metric. It does not score as highly in Precision as it indiscriminately pulls nouns and verbs from the text and ends up with many extra words that are unused in the reference summary. This baseline system also purposefully does not reference a dictionary created or updated for the CRD3. One of the improvements we focused on with our own summarization system was ensuring it could handle the unique dialogue and vocabulary present in the corpus, and this baseline allows us to compare to a system that does not factor that in. As a part of that, the baseline system merely skips any OOV words it finds, which expands and better highlights the difference between a naive system and a more refined system.

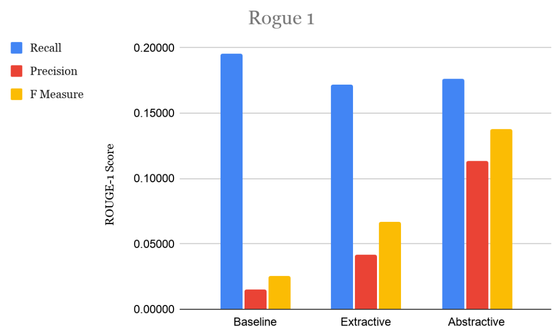
7.4 Test Set

The CRD3 includes 115 episodes from the first season of the show, which served as the training corpus for our system. The CRD3 also includes 46 episodes from the ongoing second season of the show. We created our test corpus by selecting a random sample of 5 episodes from this second season and used the remaining 41 episodes as our development corpus. One challenge in getting our results was the runtime of our final extractive summarization system. With each episode summary taking approximately forty minutes to generate, assessing our system throughout the process became a bit of a challenge.

8 Results and Discussion

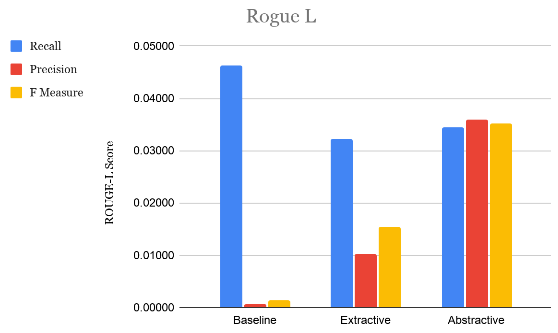
Our final results display the average Recall, Precision, and FMeasure scores for our system using the ROUGE-1 and ROUGE-L assessments plotted next to the same results from our baseline system and our reference point pre-trained abstractive system.

ROUGE-1	Recall	Precision	FMeasure
Baseline	0.19533	0.01519	0.02553
Extractive	0.17180	0.04190	0.06692
Abstractive	0.17647	0.11321	0.13793



ROUGE-1 Results on our Text Corpus (Above)

ROUGE-L	Recall	Precision	FMeasure
Baseline	0.04621	0.00073	0.00140
Extractive	0.03220	0.01030	0.01549
Abstractive	0.03448	0.03587	0.035165



ROUGE-L Results on our Text Corpus (Above)

These results generated from summaries from our test corpus show that our final system did surpass our baseline system by a significant margin. The phenomenon described above where summarization systems in general do not score highly on a zero to one FMeasure score can also be observed. Illustrating this, the average test set ROUGE-1 FMeasure score for our final system is about .07. On a zero to one scale, this is not a great result, but when compared to the pre-trained abstractive summarization system's average ROUGE-1 FMeasure score of .138 it is not as low of a result. Remapping our average score we can find that our average ROUGE-1 FMeasure score on a 0 to 0.138 scale is about .485.

We also find that this average FMeasure score significantly surpasses the Baseline ROUGE-1 FMeasure score on a 0 to .138 scale of .185. We can therefore conclude that on a word/unigram level, our system shows significant improvements over a naive baseline implementation.

In terms of ROUGE-L, our average FMeasure score is .0155 on a zero to one scale or .443 on a 0 to .035 scale - the ROUGE-L score of the pre-trained abstractive system. Compared to the baseline ROUGE-L FMeasure score of .0014 on a zero to one scale or a remapped score of .04, we observe an even greater increase in performance of our system over the baseline. Therefore, we can conclude that our final system has a far superior sentence level summary performance than a naive system.

The ROUGE-L findings are additionally significant because of the nature of summarization. While matching unigrams is a marker of an effective summary, one of our primary goals for our system was to generate readable output that could be useful to a human user. The higher ROUGE-L performance of our systems indicates some level of success in generating summaries that make sense on a sentence level.

For both ROUGE-1 and ROUGE-L, we can observe that our Recall scores are very much comparable to the T-5 abstractive model. It is in terms of Precision where the extractive implementation results in a much more unnecessarily long summary, making that part of the score suffer. Given more time, we think exploring methodologies to improve specifically the precision of our system would be a good next step.

9 Future Work

One of our original ideas for use with the CRD3 is a text generation system, where a potential dungeon/game master could input key terms, and a new summary will be outputted. This output would help the game master create their world and write their story. Since we had already set up a data processing system with this data corpus, a small trial was made to see how well this type of program worked, and to see how useful it is. The inputted key terms are turned into a vector, and the most similar sentences (using cosine similarity) are outputted. Two examples can be seen below:

input: ['music', 'performance', 'circus', 'stage', 'dance']
There's a faint bit of music as two slovenly-looking musicians in the corner are trying to work for tips with a small hat on the floor that it looks like nobody's thrown any coin into it. The music suddenly comes to a stop. Their rigid form becomes fluid to the eye as they move and shift to the music the violin now sourceless once again not seeing where this Desmond is placed but they seem to move as it picks up speed. She matches the pace of the music drawing to a crescendo and leaps and barrel-turns and climaxes with her striking a powerful pose as the lights rocket to a victorious luminescence.

input: ['trees', 'forest', 'jungle', 'plants', 'overgrown', 'grass']
Now coming across the Gravid Archipelago-- and you'd know this a little bit though you've not actually traveled to these before-- you can see this trio of islands are rather rocky and are only marked with small pockets of trees and jungle. The jungle line begins about 40 feet before you and immediately grows dense and tangled with vines root and various jungle tropical trees that choke the sky above. As you guys make your way up cresting the hill down into this small localized valley toward where the shack is with the clusters of trees and the edge of a small forest cluster is a little bit past it. There are bits of trees and clusters of them nothing that would be considered a major forest. Unable to make out any other details other than the tops of these trees and

> there's a few spots in between where you can see overgrown grass and general ground rock.

The user input defined summary generation posed a unique challenge in terms of evaluation. For the summarization system, the CRD3 included human created summaries for each episode that we could compare our system to and conduct a reasonable evaluation of the results. However, as this portion of our system was generating novel output, not contained within the CRD3, we do not have a reasonable comparison subject. We did not see this as an enormous obstacle though, as this aspect of our project is not necessarily the focus, but more of an additional outcome of the way our system works, as well as being another applicable use case for users. Therefore, even without specifically evaluating the results of our generation, we can use the positive results of our summarization system as evidence this system has a certain level of performance. This does not evaluate the method in which pieces of summaries are added together, but our previous evaluation demonstrates that at least those pieces are accurate to the corpus on a unigram level and have a certain level of fluency at a sentence level.

10 Conclusion

In conclusion, thanks to the data processing to intake this unique corpus as well as the series of improvements in the actual extractive summary algorithm itself from using word frequency, to tfidf, to cosine similarity score, we ended up with a system that far surpasses our baseline. While our system does not reach the score of a pre-trained abstractive model from T5, that system became a useful comparison point for our ROUGE evaluation. That evaluation helped us make sense of the data and allowed us to conclude that our system was fairly accurate at a word level, and relatively sensical on a sentence level. Additionally, the cosine similarity approach allowed us to adapt our summarization system into a summary generator, combining pieces of the corpus to create novel summaries and while not easy to evaluate, are backed up by the efficacy of our summarization system as a whole.

11 References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E., Gutierrez, J., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. 1–9. url: https://arxiv.org/abs/1707.02268

Cheng, J., & M, L. (2016). Neural Summarization by Extracting Sentences and Words. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1–11. doi: 10.18653/v1/P16-1046

Lin, C. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Association for Computational Linguistics,* 74-81. url:https://www.aclweb.org/anthology/W04-1013/

Mihalcea, R. (2005). Language independent extractive summarization. *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions - ACL '05*. doi:10.3115/1225753.1225766

Nallapati, R., Zhou, B., dos Santos, C., Gùlçehre, Ç., & Xiang, B. (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and

Beyond. *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 1–11. doi: 10.18653/v1/K16-1028

Rameshkumar, R., & Bailey, P. (2020). Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1–14. doi: 10.18653/v1/2020.acl-main.459

Xiao, W., & Carenini, G. (2019). Extractive Summarization of Long Documents by Combining Global and Local Context. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. doi:10.18653/v1/d19-1298