FILM-SCRIPTER: OPEN SOURCE SCRIPT GENERATOR WITH A CHAR-RNN

Claire Goeckner-Wald

Caltech
MSC 318

ABSTRACT

This research looks at the work of machine-generated short film script *Sunspring*, and attempts to recreate and validate its results. The goal is to provide an open-source film script generator. This has been done by examining the methodology outlined by Ross Goodwin, creator of the model that wrote *Sunspring*, and applying a "cumulative priming" to provide sensible plot continuity. It was found that the "cumulative priming" method was unable to create any significantly continuous plot. The repository for this code is available at

`github.com/cgoecknerwald/film-scripter`

and contains all models used for this research project. In addition, the repo contains a GUI via locally-hosted Flask server.

***Index Terms*—** char-rnns, film scripts, plot, priming strings, cumulative pre-seeding

1. INTRODUCTION

As more viewers spend increasing amounts online, entertainment companies are driven to produce more content. Pew Research shows that young people are spending increasing amounts of time streaming entertainment content, and decreasing amounts of time watching traditional television [1]. Many entertainment providers, such as HBO, Hulu, Amazon Video and Netflix, offer television content to stream online, anytime. These four providers alone order large amounts of content to remain competitive with high-output free streaming services with less traditional entertainment, particularly low-budget independent creators. However, a single hour of serial television can cost between $2 and $3 million to produce [2].

Screenwriting is a mainstay of pre-production processes. High costs and slow turnaround for screenwriting can stymie productions. Because most scripts are semi-structured, and can be quantitatively analyzed, they are a target for artificial intelligence. Moreover, the same algorithm that generates the script could also assist in the identification of relevant locations, props, cast members, costumes, special effects, and visual effects. Semi-structured scripts have been shown to be quantitatively analyzable, allowing for a sufficiently intelligent algorithm to be able to recreate plot movements characteristic of episodic productions [2]. Natural language processing (NLP) has been enhanced to the point that screenwriting is now a valid option, provided we invest in long-term research. By investing in research for artificially created scripts, we can substantially enhance the entertainment industry with lower-cost, faster-time pre-production. This would allow for traditional studios, such as CBS Television Studios, to increase their productivity and throughput, in order to remain competitive with streaming services.

2. BACKGROUND

In 2016, Oscar Sharp, BAFTA-nominated filmmaker, teamed up with AI researcher Ross Goodwin to produce what some believe to be the first film with an artificially generated script [3]. The short film was produced for the 2016 48-hour Film Challenge of film festival Sci-Fi-London. The algorithm, which named itself Benjamin, was trained on science fiction scripts, which affected its output. The script was edited by the production team to fix some 'garbled script', but it is unclear what type of editing was done because the original script was not published. Furthermore, the director instructed the actors to interpret the script freely. With the added context provided by props, delivery, and direction, the final product of *Sunspring* was alternatingly "hilarious and intense". Critics called it a "novelty" and a "thought experiment", as well as "fascinatingly incoherent" (Furness, 2016; Newitz, 2016). By using the algorithm to produce the script, the production team was able to whittle down pre-production time to essentially negligible minor edits, allowing the full 48 hours to be spent on filming and post-production processes ("This Is What Happens When an AI-Written Screenplay Is Made into a Film," 2016). Thus, the algorithm expanded the working time for other, non-substitutable processes.

As demonstrated by some of *Sunspring*'s quirks, it is important that the script has context - both within itself, and to the outside "world" within the film's universe. In the screenwriting world, scripts are broken down into acts, sequences, scenes, and beats. Each term describes a unit of change within the film.

In 2009, "The Structure of Narrative: The Case of Film Scripts" attempted to quanititatively analyze film scripts. As

shown in Figure 1, Murtagh et. al. were able to quantify the script dialogue such that scenes could be clustered by relationship to the overall plot. This is important for script generation because it allows us to quantify the amount of change occurring in each scene. Therefore, we could theoretically apply some important machine learning concepts to the relatively abstract quality of plot. By applying our algorithm to the plot structure, we could theoretically rectify the quirks of *Sunspring*. The clustering of these beats represents the contribution of each beat to the total 'scene'. Beats that are clustered together are beats that are conceptually closely related. For example, beats 5, 6, and 7 are each of main characters Rick and Laszlo expressing rapprochement towards one another. Thus, they are closely conceptually related.

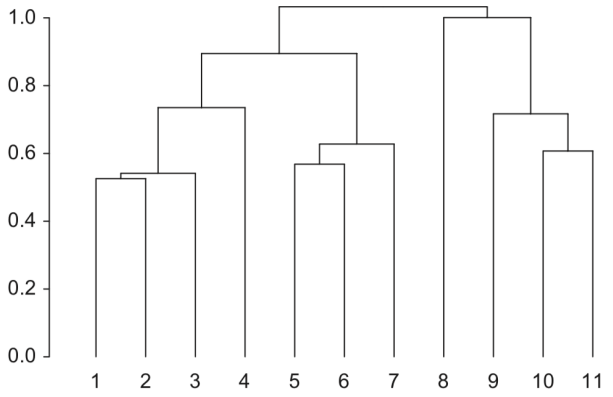

Fig. 1. Hierarchical Clustering of 11 beats from *Casablanca* [2].

However, no legal dataset appears to exists that can provide sufficient tagged and ordered dialogue to train using Murtagh's algorithm.

3. DATA

Screenplays are multi-faceted, and store a variety of data types concurrently. Most film script databases have erratic storage methods and no standardization. Hand-cleaning is not effective, since we need "25-100 MB raw text, or 50-200 novels" for this LSTM. [3] In addition, most film script databases are operating in a legal grey-zone, due to US and international copyright laws.

Instead, one can combine several different datasets.

- *Cornell Movie-Dialogues Corpus:* "conversations extracted from raw movie scripts: 220,579 conversational exchanges…involving 9,035 characters from 617 movies." [5]
- *Project Gutenberg:* "over 57,000 free eBooks…not protected by U.S. copyright law." [6]

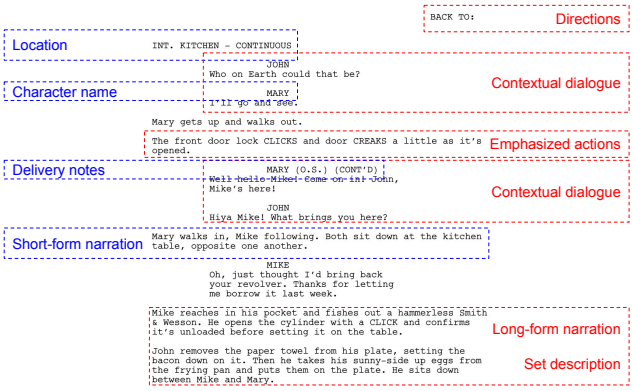

Fig. 2. Annotated example industry-standard screenplay. [4]

THREEPIO Are you sure this things safe? EXT. TATOOINE - ANCHORHEAD SETTLEMENT - POWER STATION - DAY Heat waves radiate from the dozen or so bleached white buildings. Luke pilots his Landspeeder through the dusty empty street of the tiny settlement. An old woman runs to get out of the way of the speeding vehicle, shaking her fist at Luke as he flies past. WOMAN I've told you kids to slow down! INT. POWER STATION - DAY

PROMOTER ...Balboa!?Rocky raises his head. The promoter steps over. PROMOTER (continuing) ... Twenty bucks for the locker an' cornerman -- Two bucks for the towel an' shower, seven for tax -- The house owes ya, sixty-one dollars.The man peels off the money and departs... Rocky closes hislocker, nods to the defeated fighter, and leaves.INT. TROLLEY - NIGHTRocky is on the trolley heading to South Philly... His hair lookslike it has been shaped with hedge clippers. His name isMIKE. ROCKY Yo, Mike -- What's happenin' here?

EXT. GREENBOW, ALABAMA Mrs. Gump and young Forrest walk across the street. Forrest walks stiffly next to his mother. FORREST (V.O.) Now, when I was a baby, Momma named me after the great Civil War hero, General Nathan Bedford Forrest... EXT. RURAL ALABAMA A black and white photo of General Nathan Bedford Forrest. The General is in full Ku Klux Klan garb, including his horse. FORREST (V.O.) She said we was related to him in some way. And, what he did was, he started up this club called the Ku Klux Klan. They'd all dress up in their robes and their bedsheets and act like a bunch of ghosts or spooks or something.

Fig. 3. Example scraped screenplays: *Rocky*, *Forrest Gump*, and *Star Wars: A New Hope*, respectively.

- *CMU Movie Summary Corpus:* "42,306 movie plot summaries extracted from Wikipedia…" [7]

4. PLOT STRUCTURE

Traditionally, the *plot* features a *climax* and has 2 - 3 *acts* (forming the macrostructure). Each act is composed of *sequences*; each sequences is composed of *scenes*; a *beat* is a unit of action or behavior (forming the plot's microstructure).

$$\text{plot} \supset \text{acts} \supset \text{sequences} \supset \text{scenes} \supset \text{beats}$$

5. SCRIPT-GENERATION

Given the three datasets available, it was theorized that one could use priming strings to recreate plot structure without the excessive dialogue-tagging necessitated by Murtagh's proposal. By first generating a plot summary, one could prime or "pre-seed" the output generated by the narration

model or the dialogue model, in order to confine the narration and dialogue to the scope of the plot summary.

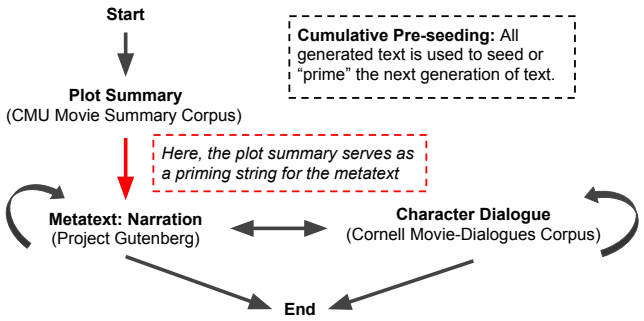

Fig. 4. Proposed script generation via cumulative pre-seeding

In theory, this could be expanded. One could generate a plot, manually section it into the traditional 3 acts, then train narration and dialogue on their respective act's plot summaries.

6. CHAR-RNN

The model was built with PyTorch. The base of the model was borrowed from Sean Robertson's `char-rnn.pytorch`, which in turn was borrowed from the Practical PyTorch series. [8] Training each model took 11 - 16 hours, using Nvidia's CUDA on an Nvidia GTX 1070. The following training stats were applied:

- LSTM (Long Short-Term Memory);
- 2^{16} (65,536) iterations;
- 2 or 3 layers;
- 512 hidden states;
- 0.01 learning rate;
- 256-char sequence length;
- 128 batch size;
- 0.25 dropout.

No significant difference was noticed between 2 or 3 layers, besides increased training time.

7. MODEL PERFORMANCE

The following sample dialogue demonstrates the abilities of the dialogue generation trained on the Cornell Dialogue Corpus. The model was primed with "Where", and demonstrates increasing temperatures (Newlines: `\n`)

Temperature	Output
`0.2`	Where are you going to do to you? `\n` What are you doing to see the street? `\n` What do you mean? What are you talking to you? `\n` What are you doing?
`0.4`	Where are you so change? `\n` What do you think? `\n` So what are you talking to you a lot of beat the British of you are going to do to you?
`0.6`	Where are you talking to you to see the same? `\n` What did you get his starship? `\n` What do you want out? `\n` Oh, no. I like you.
`0.8`	Where are you? `\n` Yeah, I could say nothing. `\n` What? `\n` Spercoursex, sure. I saw you the time you got hype. `\n` What is your pattern and do something? You know that, someone up? Saint suicides Lieutenable, Mother. Frank.
`1.0`	Where are, bene? `\n` I thought he love afters six man with degene, Jertain... I saw me. `\n` I dunnoch - his direct travel, to see you. `\n` So–

8. CONCLUSION

There is a lack of comprehensive film script training sets. This research provides open-source tools to work around erratic literature from Project Gutenberg (see `gutenberg_parser.py` and others). Furthermore, Char-RNNs, though fast learners, are too erractic at high temperatures. Finally, simple priming strings are ineffective for contextual generations, and further research is needed.

The repository for this code is available at

`github.com/cgoecknerwald/film-scripter`

and contains all models used for this research project. In addition, the repo contains a GUI via locally-hosted Flask server.

9. CONTINUED WORK

- Train multitude of models on different dataset combinations (e.g., having dialogue trained on Cornell Corpus *and* Question-Answer datasets).
- Switch from Char-RNN to Word2Vec-RNN. Possibly use char-RNN for character name generators.
- Subsect generated plot summary into acts. Use the plot summary of each act to prime the generated text for that act only, rather than generating all text from the entire plot summary.

10. ACKNOWLEDGEMENTS

The author wishes to thank Ross Goodwin at Interactive Telecommunications Program NYU and BAFTA-nominated filmmaker Oscar Sharp for inspiration via *Sunspring*; Sean Robertson (`spro`) of Prontotype for the MIT-licensed basis for the char-RNN repository in PyTorch; Cristian Danescu-Niculescu-Mizil of the Department of Information Science, Cornell University, for the Cornell Movie-Dialogs Corpus; Bamman, O'Connor, and Smith at the Language Technologies Institute and Machine Learning Department at Carnegie Mellon University for the CMU Movie Summary Corpus; and the countless authors whose works in the public domain were invaluable to the meta-text training set (via the Gutenberg Project).

11. REFERENCES

[1] Lee Rainie, "About 6 in 10 Young Adults in U.S. Primarily Use Online Streaming to Watch TV," 2017.

[2] Adam Ganz Murtagh, Fionn and Stewart McKie, "The Structure of Narrative: The Case of Film Scripts," *Pattern Recognition*, vol. 42, no. 2, pp. 302–312, 2009.

[3] Ross Goodwin, "Adventures in Narrated Reality: New forms & interfaces for written language, enabled by machine intelligence," 2016.

[4] Mendaliv, "Sample from a screenplay, showing dialogue and action descriptions," https://commons.wikimedia.org/w/index.php?curid=4970002, 2008, Online; accessed December 2018.

[5] Cristian Danescu-Niculescu-Mizil and Lillian Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[6] Michael Hart, "Project Gutenberg: Free eBooks," http://www.gutenberg.org/, 1971, Online; accessed August - December 2018.

[7] Brendan O'Connor David Bamman and Noah A. Smith, "Learning Latent Personas of Film Characters," in *Association for Computational Linguistics 2013: Sofia, Bulgaria*, 2013.

[8] Sean Robertson, "char-rnn.pytorch," https://github.com/spro/char-rnn.pytorch, 2018, GitHub repository; accessed December 2018.