Developed Film: A Computational Approach to Movie Plot Generation

Alex Bullard
Middlebury College
Middlebury, VT
abullrd@gmail.com

William Martin
Middlebury College
Middlebury, VT
will.c.mrtn@gmail.com

Joe Redmon
Middlebury College
Middlebury, VT
pjreddie@gmail.com

Abstract

Automated plot generation is of great interest to the film industry, which relies heavily on novel and innovative stories. Current approaches require an extensive number of plots broken down into labeled components, such as character relationships or actions. Unfortunately, this does not always result in the most earth-shattering stories and necessitates a human to manually write details about the data. We present an approach that does not require labeled data, but instead creates unique stories from the statistical analysis of a massive data set. Utilizing a combination of probabilistic context-free grammars and *n*-gram models based on our data, we generate plots that are in some cases, both grammatically correct and captivating.

1 Introduction

Humans are natural storytellers with the ability to craft a wide range of narratives spanning numerous topics and this ability to create plots that are grammatically correct and semantically sensible is often taken for granted. Within the field of Natural Language Processing story generation is an extremely interesting topic because it is very difficult. Generating believable stories requires a sense of plot, correct syntax and semantics and continuity. Although there have been several attempts to generate stories, computers are still unable to create plots indistinguishable from those written by humans. This research attempts to solve the more feasible sub problem of how to computationally generate movie plots.

Automated plot generation has a number of potential advantages over manual story crafting. Stories are traditionally a static experience; why not make them dynamic and unpredictable? Audiences crave an immersive experience, which can be magnified by allowing plots to react to a user's wishes. Similar to "choose your own adventure" novels, computational plot lines could allow for an interactive experience in any medium that requires a story. In addition, this would allow various media to be tailored towards specific demographics. The advent of computationally generated plots would allow for a significantly more captivating experience for story consumers and lower content costs for producers.

Previous work on how to automatically generate stories has yielded a number of varied results. Gervás et al. (2005) presented a case-based reasoning algorithm for generating plot structures from existing stories. Although this method was able to generate sensible broad plot outlines, it required hand crafted details about the initial stories and gave very general sentences. A similar generation technique was used by Young (2007) in a bipartite model that separated plot events (story) from how they are presented (discourse). However, this algorithm again is best at generating plot plans, which do not allow for explicit details. One of the most fluid story generation algorithms is given by McIntyre and Lapata (2009) and focuses on statistically analyzing raw story data. This allows the authors to generate plau-

sible short stories that are organic and grammatically correct. This paper focuses on a similar approach with a focus on allowing more diverse grammars and more complex training data.

2 Algorithm Description

Our algorithm divides the process of movie plot generation into four distinct stages as shown in figure 1. We begin by preprocessing raw data from the Internet Movie Database (IMDb) and use it to generate multiple models that describe the structure of movie plots, titles, characters, and taglines. These models utilize both n-grams and probabilistic context-free grammars (PCFG) in order to randomly generate new unique movies.

During the preprocessing stage various plain text documents from IMDb, including actors, genres, plots, etc., are conglomerated into a single file. Plots are further processed with the use of IMDb's actor database in order to standardize character and actor name references across sentences and plot summaries. Special care was taken to ensure that the processed data were as high quality as possible. Because IMDb's plot data is submitted by users and not by paid staff, its quality is frequently spotty and includes inconsistencies like misspelled character and actor data. This was mitigated through the use of various techniques, including the use of common nickname lists and partial name matches.

Sentence generation is accomplished with the following steps. First we generate a PCFG for each of the genres from our input plot summaries; these represent the form of plot sentences within their respective genres. For a requested plot generation, we use the appropriate PCFGs to create some number of initial sentence forms. In addition, we create a n-gram model along with a higher order set of grams for each of the genres. Using each of the n-gram models, we generate a large number, on the order of thousands, of sentences. Within these sets of sentences, we use the high order sets of grams as "out" lists to select sentences for removal. This ensures that we do not use any plot snippets that too directly match our original input data. From what remains of each sentence set, we select the sentences that most closely match the original PCFG sentence forms based on an F1 score. These are the sentences that finally make up our plot.

Taglines and titles are handled with a bigram model due to their relatively small size. Any n-gram model higher than this grossly overfits to our input data. We also select actors and actresses for our films based on the placeholders we initially added to our input plots. They are chosen randomly from a list of high-grossing, popular actors so that they are well recognized by the reader.

3 Results

In order to evaluate the performance of our plot generation algorithm, we conducted a series of experiments that measured the plausibility of both plot summaries and individual sentences. We presented our subjects with 50 plot summaries and 50 sentences that consisted of both computer generated and human written samples. Test subjects were then instructed to give two grades to each sample from 1 (worst) to 5 (best) on grammar and semantics. Plot summaries were given an additional grade on how well organized they were. Finally, they answered whether they believed each sample was computer or human generated. This allowed us to determine how our computationally generated sentences compared to those created by humans.

Our results show that there is still a large gap between the content we generate and the content produced by humans. The average discrepancy of 1.62 between the rated syntax of the sentences is probably a result of the fact that sentences are most reliant on n-gram generation which does not use grammar rules. Semantically, individual sentences did surprisingly well, with the much lower score difference of 1.07 and an average of 3.14. Unfortunately, as might be expected, the plausibility of our full generated plots was much lower. Since the only connection between sentences is grammar related and does not use context, the plots have a tendency to include non sequitur and none of the plots were rated as plausible.

4 Conclusion

This paper presents a statistical approach to generating random movie plots, titles, and taglines. Although the evaluation of the generated sentences proves that on average they are not as close to real

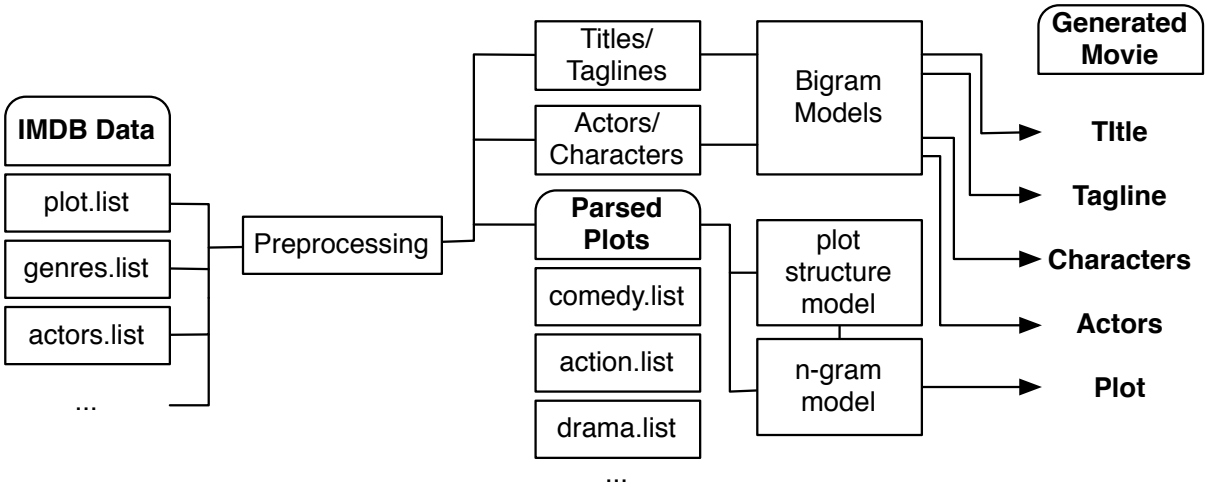

Figure 1: Block diagram of program flow.

Table 1: Evaluation results

Test	Result
Sentences	
Human - Grammar	4.57
Computer - Grammar	1.63
Human - Semantics	4.21
Computer - Semantics	1.07
Human - Correctly Identified	0.36
Computer - Correctly Identified	0.78
Plots	
Human - Grammar	4.80
Computer - Grammar	4.20
Human - Semantics	4.60
Computer - Semantics	3.40
Human - Correctly Identified	0.40
Computer - Correctly Identified	1.00

In a Future not so far future the world has suffered an unspecified catastrophe.
James arrives home to find his place in the world of organised crime.
Determined to get back the money, he decides to take his revenge.
His mission leads him to the heart of the heavily armed enemy.

Figure 2: Sample generated sentences

sentences as might be desired, individual sentences are sometimes surprisingly believable. The most poignant realization made during this research was just how difficult of a problem natural language generation is. One must understand the plot structure of story in order to generate its final discourse. Our algorithm attempts to imitate this, but it is not a true creative process. Thus, we often generate sentences that do not progress story events smoothly or sensibly.

The ultimate goal of this research is to automatically generate plot lines that are indistinguishable from their human-written counterparts. There are a number of possible improvements that could be made towards this end. Additional data processing and gathering could significantly improve our final

results by increasing both the breadth and quality of our training stories. Furthermore, by modifying our models to further distinguish between discourse and story, we could improve our plot generation by encouraging coherent across sentence boundaries.

References

Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás 2005. Story Plot Generation based on CBR. *Applications and Innovations in Intelligent Systems XII*, 33–46.

R. Michael Young. 2007. Story and discourse: A bipartite model of narrative generation in virtual worlds. *Interaction Studies*, 8(2):177–208.

Neil McIntyre and Mirella Lapata. 2009. Learning to tell tales: a data-driven approach to story generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 217–225.

David D. McDonald. 1992. Natural-Language Generation. In *Encyclopedia of Artificial Intelligence*, 642–655.