Genre Classification with Video Game Transcripts and Descriptions

Sayer Rippey and Ethan Zimmerman

Natural Language Processing

1. Introduction

This paper reports on an experiment which applied the Natural Language Processing problem of genre classification to video game descriptions and transcripts. More specifically, given a corpus of video game descriptions and a parallel corpus of transcripts, we investigated the feasibility of automatically classifying video games across a set of not mutually exclusive video game genres, including Fantasy and Horror as well as others (first using the transcripts, then the descriptions, then both). Because different genre labels categorize video games in various ways, we anticipated that different video game genres would be classified with varying degrees of difficulty, as each genre might be expressed in radically different types of linguistic cues. For that reason, and because scripts are longer (and so contain more information), we also hypothesized that the script corpus would generally be more useful than the description corpus, but not necessarily for every game genre.

This work has potential implications that are both practical and academic. Automated genre classification could be useful in game recommender systems, and applying NLP to video game transcripts more broadly could be useful to video game developers by identifying linguistic trends in various types of video games, such as more/less successful video games, or video games targeted at different demographics. NLP work on video games is more generally important in the study of video games and their interactions with culture. Video games are now an incredibly prolific medium, with 97% of teenagers playing video games in the US [5]. Researchers have explored their potential uses in education [3] or training [6], as well as their behavioral and social effects [1]. NLP techniques can reveal linguistic patterns that have implications for all of these domains. Also, many other fields are pursuing research into video games and their effects. A successful video game classifier would ease some of these researchers' time if they hope to study a specific genre of games.

2. Related Work

To our knowledge, no one has applied NLP techniques to video game transcripts or descriptions, making this a potentially very exciting field of work. However, there is some precedence for NLP work involving video games. Zagal et al [8] analyzed gaming reviews in order to identify their average readability in terms of grade level, as well as to come up with a list of words that are described as features of gameplay and how words with a generally negative connotation can be used in a positive sense within reviews. They used methods ranging from word and sentence difficulty analysis, POS tagging, and sentiment analysis. With their experiments, they hoped to give gamers and game designers a better framework within which to analyze games and game reviews.

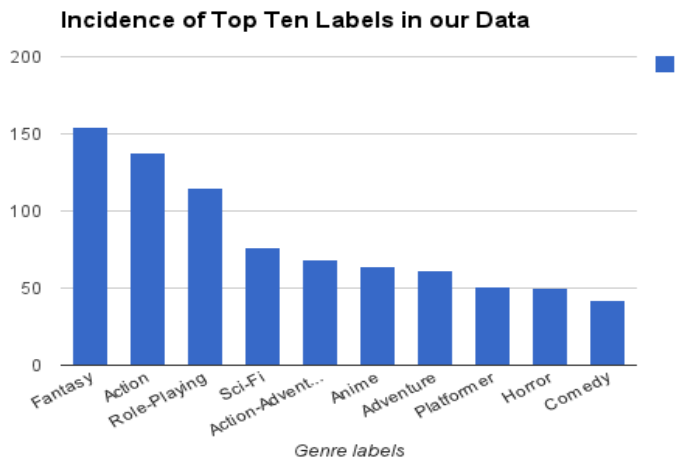
Genre classification has also been explored in the NLP community. Blackstock and Spitz [2] used Naive Bayes and a Maximum Entropy Markov Model to classify movie genres from movie scripts. This is very similar to our work in that movie genres are not mutually exclusive, although with some potentially

significant differences. One difference is that video game genres may categorize video games in more ways than movie genres categorize movies (eg, video game genres might define video games according to the perspective of the player, the goal of the video game, the primary activities in the video game, or etc). Another possible difference is that the information included in a video game transcript may be very different from the information in a movie script, meaning different types of features may be useful.

3. Data

For our project, we obtained a corpus of video game transcripts from VGScripts.com [7]. There are 319 video game scripts, which were transcribed or uploaded by VGScripts users. We used Scrapy to get the transcripts. These constitute our first corpus. We then used the Giant Bomb API [4], an API which provides access to Giant Bomb's large, structured database of video game information, to get a description and a list of genres for each of the video games from our first corpus. The descriptions constitute our second, parallel corpus. The genres are the set of non-mutually exclusive labels that we were trying to classify.

There were 49 genres total, but we only experimented on the ten that appeared most frequently in our corpus, as shown in the graph below. After these ten, the counts dropped into the twenties or less.



This is not a very large corpus, as machine learning NLP experiments go, but we were limited by the number of video game scripts available to us.

4. Experiments

Our task was to build a classifier that could learn a set of labels (video game genres) on video game transcripts, video game descriptions, and on both at once. Because our labels aren't mutually exclusive, we trained our classifier separately for each label, and then performed binary classification for each genre for each video game (in three different experiments - one on each corpus, and one on both at once). This allows us to see if certain types of genres are more detectable than others, while

also investigating whether transcripts or descriptions provide more important contextual information.

We used NLTK's Naive Bayes classifier for this experiment, and feature vectors that consisted of bag of words and bigram collocations (we were originally going to use other features, but initial experiments showed that they weren't useful). We split our data into a random training set (169 games) and test set (150 games) for each experiment. The metrics we use to determine the success of the experiments are precision, recall, and f-measure. We use a guessing baseline, that is, we would like to have an f-measure higher than .5 at the very least. (Since no one has done this kind of work before, we can't really compare our work against others'). (If we get a worse f-measure, that may imply that the answer to "is this work feasible" is "no").

5. Results

We found no significant difference in overall performance between scripts, descriptions, and the combined corpus. However, for the classes that appeared less in our data, combining our descriptions and scripts yielded generally higher f-measures. However, it also decreased the f-measures for the top three labels (Fantasy, Action, and Role-Playing), probably because these already had enough data, and the extra data just contributed noise. The table below shows the accuracy, precision, recall, and f-measure across all data sets and classifying labels.

	Fantasy		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.80000	0.76667	0.56667
precision	0.74468	0.70270	0.52555
recall	0.92105	0.97500	1.00000
f-measure	0.82353	0.81675	0.68900

	Action		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.62000	0.59333	0.57333
precision	0.83333	0.66667	1.00000
recall	0.08197	0.06349	0.04478
f-measure	0.14925	0.11594	0.08571

	Role-Playing		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.66000	0.39333	0.38667
precision	0.48454	0.38514	0.38667
recall	0.97917	1.00000	1.00000
f-measure	0.64828	0.55610	0.55769

	Anime		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.69333	0.17333	0.20667
precision	0.29412	0.15646	0.20667
recall	0.31250	1.00000	1.00000
f-measure	0.30303	0.27059	0.34254

	Adventure		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.77333	0.18667	0.30667
precision	0.35000	0.18667	0.18750
recall	0.25000	1.00000	1.00000
f-measure	0.29167	0.31461	0.31579

	Platformer		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.79333	0.87333	0.86667
precision	0.40476	0.75000	0.50000
recall	0.73913	0.14286	0.05000
f-measure	0.52308	0.24000	0.09091

	Sci-Fi		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.28667	0.59333	0.56667
precision	0.28378	0.31646	0.42202
recall	0.97674	0.78125	0.95833
f-measure	0.43979	0.45045	0.58599

	Action-Adventure		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.20000	0.68667	0.60000
precision	0.18493	0.35849	0.32895
recall	0.96429	0.59375	0.73529
f-measure	0.31034	0.44706	0.45455

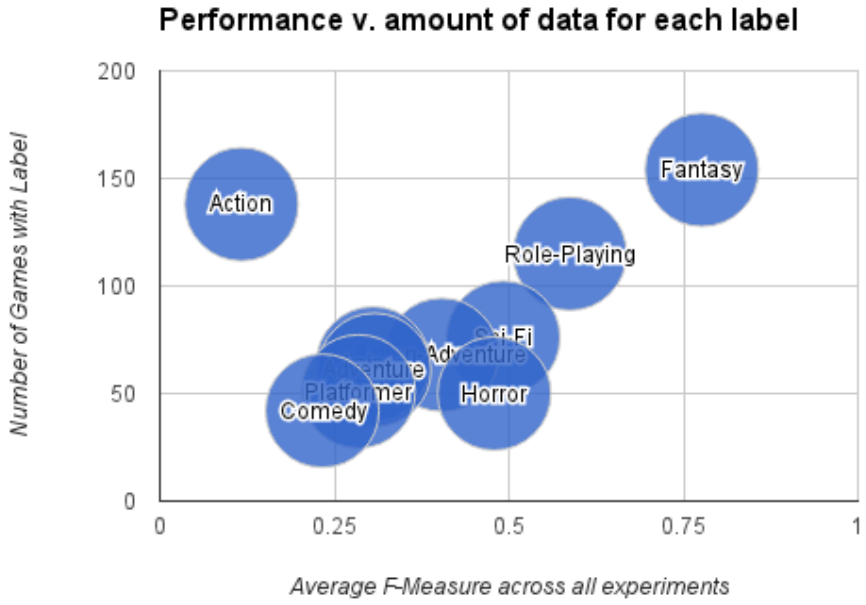
	Horror		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.20667	0.93333	0.80000
precision	0.19048	0.75000	0.42553
recall	1.00000	0.42857	0.86957
f-measure	0.32000	0.54545	0.57143

	Comedy		
	DESCRIPTIONS	SCRIPTS	BOTH
accuracy	0.12000	0.10667	0.54667
precision	0.10811	0.10067	0.19512
recall	1.00000	1.00000	0.88889
f-measure	0.19512	0.18293	0.32000

The graph below shows the average f-measure over all 3 experiments for each genre as a function of the related scripts. While there are some outliers, the general trend is a higher f-measure is associated with a greater number of scripts in the corpora. This outcome suggests that we would have achieved higher f-measures overall if our corpora had contained more scripts. The greatest outlier classification is the action game tag, which performs poorly in all three tests. A possible reason for this

is that action is a large and diverse genre; some action games have long, in-depth scripts and others have short scripts without dialogue. The poor performance of this tag suggests that different features may perform better on these games than the ones we included.

Horror was another outlier, doing significantly better than other labels with similar amounts of data. We hypothesize that this is because horror video games have very distinctive language (as corroborated by our important features, below), and because "horror" as a genre is a lot less vague than "action" or "comedy", both of which could encompass many different types of games.



The important features for each experiment are shown in the table below. There are a few interesting things happening here. One is that while the descriptions contained less data than the scripts, they frequently contained the name of the genre we were trying to classify, which was obviously a very important feature.

FANTASY		ACTION		ROLE-PLAYING		SCI-FI		ACTION-ADVENTURE	
SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTION
begone	magical	master's	while	inn	towns	accessing	occupied	deflect	german
you...you	fantasy	mt	princess	warp	shops	broadcast	mech	donald	appearing
festival	turn-based	shard	spells	bandits	spells	manually	movements	glides	cat
whaddaya	warrior	resides	statue	heals	role-playing	marines	galaxy	olympus	throughout
despise	unleash	desolate	elemental	grandmother	restored	planetary	technology	colossus	href
banish	elemental	sleepy	pretty	trustworthy	random	enhancements	cruiser	('ascii', 'art')	9.0c
lass	mechanics	starve	saw	somewhere = None	kingdom	schematics	junkyard	|_	scheduled
pawns	ammo	students	titled	banished	hp	transmitter	helmet	dracula	equally

ANIME		ADVENTURE		PLATFORMER		HORROR		COMEDY	
SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTIONS	SCRIPTS	DESCRIPTION
dividing	declared	observatory	('adventure', 'game')	cyber	platformer	dormitory	horror	simpson	joe
all-powerful	data-ref-id="3045-2	fatherly	writer	zero's	darker	trapdoor	mutated	jamaican	baseball
no...not	('real', 'time')	vine	('well', 'as')	maverick	rockman	dangles	translated	beaver	city
oww	('that', 'the')	chilled	('other', 'games')	("let's", "go")	mavericks	gauze	zombies	sato	cash
horned	level	gum	href	bed	lasers	blurred	href="/konami-"	whiskers	blocked
then...what	ranges	nests	compelling	2008	tails	scribbles	reincarnation	masato	car
oven	('has', 'a')	new-found	('can', 'also')	down	('on', 'his')	agape	respawn	chunky	stab
wrinkles	gamespot	loosely	light	whoever	halt	107	struggles	reclining	urban

We can also see that many of the words are specific to games in the script corpus, as opposed to generally describing the gameplay or experience. For example, “Donald” in Action-Adventure and “Zero’s” in platformer refer to characters in games with these tags. This could be caused by including many members of the same series in our already small corpora. This just reinforces that a larger corpus would make our results more meaningful and generalizable.

The table also included certain symbols or phrases that came with the html or script context surrounding the desired data. A possible solution to this would be to clean out the files by hand before beginning experiments.

The Sci-fi column stands out as having related word features over in the description and script. All of the words in both lists appear to be words that would appear in a Sci-fi story. While none of the words overlap and, on the base level, this is not statistically significant, in an experiment that considered semantics and schemas these features would be considerably more helpful.

On the other hand, the words that appear as the Anime features for both the description and scripts are fairly poor definitions of the genre. In the script group, some words are exclamations while others are meaningless phrase fragments or overall random words. This could very well be a factor of the few scripts we have. It could also be caused by incredibly long scripts or by very general unexplainable plots. The description side of the Anime feature words consist mainly of common word phrases, general gamer related information, or random html text. These all could be beneficial considerations for future work.

6. Conclusion

Our two hypotheses were that, 1) different video game genres would be classified with varying degrees of difficulty, and 2) that the script corpus would generally be more useful than the description corpus. Although the success of each different genre classification was largely dependent on the amount of data we had for it, we also saw that this wasn’t entirely true, and discussed some of the outliers. Future work here could include exploring that, with more features, and controlling for the

amount of positive examples of each genre.

Our second hypothesis did not turn out to be true, probably because the high level, descriptive information provided by the description corpus balanced out the script corpus' advantage in terms of amount of data.

Overall, our experiments yielded some interesting observations, although the limitations of our data prevent us from saying anything conclusive about the potential for genre ID in game scripts. The greatest fault and limiting factor in our experiments is the general lack of scripts. With a larger corpus of scripts, we could have more successfully used our description database, included greater numbers in our training/testing split, and probably have yielded more telling results. We believe that the findings our experiments have yielded should encourage further work in the field. Before such work occurs, a good source of scripts must first be established. The next step may be a vocal recorder that saves spoken text and can be run on the vast number of games out there, in order to compile comprehensive and textually clean scripts. We hope that such future work will allow for more endeavors investigating video games and Natural Language Processing.

References

1. Anderson, Craig A., and Brad J. Bushman. "Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-Analytic Review of the Scientific Literature." *Psychological Science* 12.5 (2001): 353-59. Web.

2. Blackstock, Alex, and Matt Spitz. "Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features." *Stanford NLP* (2008): n. pag. Print.

3. Gee, James Paul. "What Video Games Have to Teach Us about Learning and Literacy." *Computers in Entertainment* 1.1 (2003): 20. Web.

4. *Giant Bomb API*. N.p., n.d. Web. 10 Dec. 2013. <http://www.giantbomb.com/api/>.

5. Lenhart, Amanda, Joseph Kahne, Ellen Middaugh, Alexandra Macgill, Chris Evans, and Jessica Vitak. "Teens, Video Games, and Civics." *Pew Internet & American Life Project*. N.p., 16 Sept. 2008. Web. 10 Dec. 2013. <http://www.pewinternet.org/Reports/2008/Teens-Video-Games-and-Civics/07-14-Parents-an

d-Games/02-A-majority-of-parents-are-aware-that-their-children-play-video-games.aspx>.

6. Rosser, J. C., P. J. Lynch, L. Cuddihy, D. A. Gentile, J. Klonsky, and R. Merrell. "The Impact of Video Games on Training Surgeons in the 21st Century." *Archives of Surgery* 142.2 (2007): 181-86. Web.

7. *VGScripts.com*. N.p., n.d. Web. 10 Dec. 2013. <http://www.vgscripts.com/>.

8. Zagal, José P., Noriko Tomuro, and Andriy Shepitsen. "Natural Language Processing in Game Studies Research: An Overview." *Simulation & Gaming* 43.3 (2012): 356-373.

We have adhered to the Honor Code on this assignment. -Ethan Zimmermann, Sayer Rippey