AUTOMATIC GENERATION OF TEXT FOR MATCH RECAPS USING ESPORT CASTER COMMENTARIES

Oluseyi Olarewaju[1], Athanasios V. Kokkinakis[1], Simon Demediuk[2], Justus Roberstson[1], Isabelle Nölle[1], Sagarika Patra[1], Daniel Slawson[1], Alan P. Chitayat[1], Alistair Coates[1], Ben Kirman[2], Anders Drachen[2], Marian Ursu[2], Florian Block[2] and Jonathan Hook[2]

[1]Weavr, Arena Research Cluster (ARC), University of York, York, UK
[2]DC Labs, Arena Research Cluster (ARC), University of York, York, UK

ABSTRACT

Unlike traditional physical sports, Esport games are played using wholly digital platforms. As a consequence, there exists rich data (in-game, audio and video) about the events that take place in matches. These data offer viable linguistic resources for generating comprehensible text descriptions of matches, which could, be used as the basis of novel text-based spectator experiences. We present a study that investigates if users perceive text generated by the NLG system as an accurate recap of highlight moments. We also explore how the text generated supported viewer understanding of highlight moments in two scenarios: i) text as an alternative way to spectate a match, instead of viewing the main broadcast; and ii) text as an additional information resource to be consumed while viewing the main broadcast. Our study provided insights on the implications of the presentation strategies for use of text in recapping highlight moments to Dota 2 spectators.

KEYWORDS

Esport, Data-Driven Storytelling, Dota 2, Game Analytics, Broadcasting, social viewing, Linguistic Resources, Natural Language Generation.

1. INTRODUCTION

Advances in Natural language generation (NLG) and data-to-text tools have found relative success outside esport (e.g. weather forecasting, robojournalism, traditional sport summaries and fantasy football). Bringing the richness of automated storytelling to audiences of esport using NLG and data-to-text techniques by generating text-based recaps creates a unique selling point for content producers who may want to appeal to a wider audience beyond expert esport players and viewers. Automatically generated text can also be easily customized to reflect preferences that are particular to players' style and level of experience of the viewers. This short form of highlight personalization can help gaming operators maintain an active player base who otherwise may not have the time or resource to commit to viewing an entire esport game. In addition to the proposed support by [1] for amateur casters such as pairing casters for live games in a matchmaking queue. Automatically generated text recap offer opportunity to create professional level text recaps similar to those that are written by humans for lower leagues that have fewer resources.

David C. Wyld et al. (Eds): CSEA, DMDBS, NSEC, NETWORKS, Fuzzy, NATL, SIGEM - 2020
pp. 117-131, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101810

One of the most significant differences between esports and traditional sports is the level of complexity of the gameplay. Complexity has been known to create significant barriers to esport enjoyment and may take away the initial appeal that the aesthetic aspect of the games may have made to players/spectator [2]. As esports popularity grows, content creators need to present the complexities in the games played in understandable ways without comprising on the quality of insights they provide. This motivates the research presented here, we study the use of text-based recaps generated by Natural Language Generation (NLG) systems for presenting highlight moments in esport games (e.g. Dota 2). Can content producers communicate non-linguistic esports data to better engage their audience and make watching esports more meaningful and enjoyable experience? While the potential of these approaches is enticing, it is not yet understood whether current state of the art techniques for NLG can be used to create text summaries the are accurate and engaging for audiences and what value these summaries can have for audiences during and around the spectating of matches. Our work aims to contribute to knowledge in this area by addressing these questions using a generation grammar derived from parsed live commentaries to automatically generate text-based recaps of most important actions and highlights of Dota 2 games.

Text-based recaps are translations of non-linguistic performance data of players in a match to linguistic data stories. Using production rules from the generation grammar, inputs obtained from game data are mapped to parsed commentaries to generate recaps for moments of highlights in matches. On one hand, we investigate if spectators perceive automatically generated text from the grammar as accurate recaps of the highlight moments in Dota 2 matches. On the other hand, we asked if text recaps generated by the derived grammar support understanding of highlight moments in Dota 2 matches by considering two use cases. In the first case, we considered the potential use of automatically generated text as an alternative way to spectate a match, instead of viewing the main broadcast. In the second case we considered a combination of linguistic and visual information for recapping highlight moments and investigate using text as an additional information resource to be consumed while viewing the main broadcast.

In many real life applications, it is essential that text outputs from Natural Language Generation (NLG) systems accurately inform end users about input data that is being communicated. According to [3], hallucination and other forms of inaccuracy are unacceptable in NLG application contexts such as journalism, financial reporting, and medical patient information. Esports, although an emerging industry is without exception. User analytic are important for improving skill level and all non-linguistic data need to be communicated correctly. In this work we used handcrafted features with grammar rules derived from commentary corpus to support the linguistic communication of user performance data in easy-to-interpret text formats. Commentaries in esport games are one of the ways that enjoyment is sustained by casters who provide a hyped narrative of play or events of the game. Entertaining and informing audiences about different aspects of a live game are two of the roles casters have to undertake. To keep audiences engaged, the comments that casters make during games are often articulate and colourful. These characteristics of esport commentaries make them an interesting resource for inducing rule-based grammars that can be used to generate text-based recaps which closely resemble human-written summaries.

In Section 2 we discuss previous attempts to using text corpus for generating text. We describe our approach to deriving a generation grammar and detail the process of generating text recaps in Section 3. In section 4, we provide details of our lexical resource and the corpus creation process. In Sections 5 and 6 we present our study design and the results of evaluations of the NLG System. The implications of our results for esport domain is discussed in section 7. Finally, we present our conclusion and future ideas in section 8.

2. RELATED WORK

According to [2] and [4], Esports are video games played by professional gamers that are broadcast online. Esport attracted a global viewership of 454 million in 2019 and has been projected to have a year on year growth of 15% over the next four years [5]. Esports popularity has also drawn the attention of national governments as potential means of job creation and revenue generation and has prompted large government research into exploring Esports as a new immersive data-driven experience. A point in case is the Weavr project in [6] which is government-sponsored consortium project bringing together partners with extensive expertise in a range of field relating to the creation of data-driven experiences. In traditional sports such as football, data-driven contents have been shown to enable content producers in reporting and explaining things that were not previously explicable [7]. Research in content production in Esports by [8] confirmed that data-driven content can be an effective tool to make gameplay more transparent to viewers. Using simple graphical overlays of data-driven insights, they found measurable effects on the commentary and quality of Esport coverage.

The large volume of publicly available data from Esport production has enabled the extraction of insight from esport games using novel machine learning techniques where model performance is proportional to the size of the training data. A considerable number of literature have used Esport data and published work on different insights extracted from the data using different machine learning and analytical approaches. These include classification and detection of player roles using unsupervised learning techniques for more accurate analysis [9] and [10]. A hero recommendation engine using regression and nearest neighbour algorithms was proposed by [11]. Researcher in [12] reported the performance of an encounter-based algorithm in encounter detection and win probability predictions for Dota 2 games. A few research efforts have also used visual features in video streams to identify and forecast moments of highlights within games. These include the work of [13] where they used a psycho-physiological approach and the data-driven approach to construct highlight models. Esport games use distinct visual effects to highlight key moments in a game, [14] used Convoluted Neural Networks to learn filters of those visual effects for detecting highlights. Highlights were also detected by [15] using deep learning methods on a combination of game footage, facial expressions and speech data.

Natural language and narrative storytelling researchers have used techniques in computational and theoretical linguistics to gain a better understanding of player behaviour and creating novel game experiences. Gameplay was modelled by [16] as an online mediation search and used a text-based implementation to show that the search space can embed all traversals through a game world. Insights on how player behaviour affected players' collaboration in teams were extracted by [17] and [18] by analysing texts from player discussion forums and chat logs specifically for the Esport game -League of Legends. A growing number of linguistic resources have also been obtained from crowdsourcing games or by using games with a purpose (GWAP) approaches. A dataset of sentimentally classified tweets was built by [19] using GWAP and [20] used a crowdsourcing game to create language-independent sentiment lexicons. A subject of equal importance to extracting insights for enhancing the quality of gaming experience is the communication of insight to players. In game analytics, researchers have leveraged the simplicity of text to communicate insights from otherwise complex game telemetry to viewers of Esport games. To facilitate the use of their hero recommendation engine, [11] used a text auto-complete interface to present. suggestions to users. A simple text annotations was used by [8] to wrap the core message of insights on player performances derived from historic and real-time game data to esport audience.

Closely related to effective communication of insight is the task of maintaining viewership and active engagement of esport audience. One way currently explored by esport projects is to make consumption of esport content more enjoyable by providing different ways of viewing contents. Although studies such as [6] has shown that presenting viewers with cross-platform options for consuming content through mobile applications, Virtual Reality or via interactive overlays on live stream platforms can elicit engagement with esport content. Research has yet to systematically investigate what merit the contents and the different types of format in which they are delivered provide to audiences during and around the way they view esport games. For example, will spectators who are interested in viewing a particular esport match but unable to follow a live stream benefit from textual updates in a similar way to watching the event live. Extending the narrative components of the systems that deliver textual content to include NLG techniques for outputting summaries of in-game telemetry provided us with the opportunity to design experiments that mimic different viewing scenarios for viewers.

Over the last decade, there has been a sustained drive towards the improvement of representation learning for creating language models that can be used for text generation task. A growing body of research work such as [21] and [22] have shown that domain-specific natural language tasks benefit substantially from fine-tuning on massive pre-trained corpus of text. There is an agreement in these literature that the size of the training dataset required to fine-tune language models while comparatively fewer than end-to-end approaches still requires thousands or tens of thousands of examples. The data requirements of neural-based language models have constrained their usefulness to domains where labelled data are readily accessible and reduced adoption in niche real-world applications where labelled data may be limited data. Although Neural Language approaches have become more popular in current natural language research, they are typically characterized by unexplainable outcomes with a high computational cost.

In this work, we focus on text generation approaches that build grammar rules from text corpora. Our approach aligns closely with that of [23] who built surface realizers and grammar-based NLG from corpora. In addition to interpretability and traceability of the generated textual contents offered by these approaches, the need for massive labelled data and resource-dependent tools required by neural-based models were avoided. One drawback reported in the work of [23] was the huge number of mismatches of lexical items that they recorded due to the use of different corpora and general-purpose lexicons for their grammar-based NLG system. According to [23], this type of mismatch is one of the largest problems with the use of general-purpose knowledge resources created by different researchers for different purposes.

Using grammar rules derived from a corpus of professional Dota 2 commentaries and post-match analysis, we mapped non-linguistic game data to parsed commentary texts to generate recaps in Dota 2 matches. Corpus curated from commentaries have the distinct characteristic of preserving the choice of words and professional style of commentating by professionals like the group of analysts shown in figure 1(a). Consequently, the syntactic and lexical rules derived from such a corpus can potentially enhance the readability of text generated text to recap intense moment in the game such as figure 1(b).

(a) Dota 2 The International, CC BY 2.0, via Wikimedia Commons.

(b) A frame of a highlight moment in a Dota 2 match with video effect.

Figure 1: Esport Commentators and Analysts discussion interesting moments of matches

3. AUTOMATIC TEXT GENERATION

We posed the research questions in earlier sections to investigate the accuracy of generated text from a commentary corpus and how the text may support understanding of highlight moments. Central to this investigation is a rule-based NLG system that generates the text. A general overview of the system is presented in figure 2, with a text prepossessing corpus creation component, a grammar inducing component and a text generation component with surface realization engine that facilitates the text generation. Our method induces grammar in a similar way to the technique by [23] who derived Probabilistic Lexicalized Tree-Adjoining Grammars(PLTAG) from corpora and handcrafted examples.

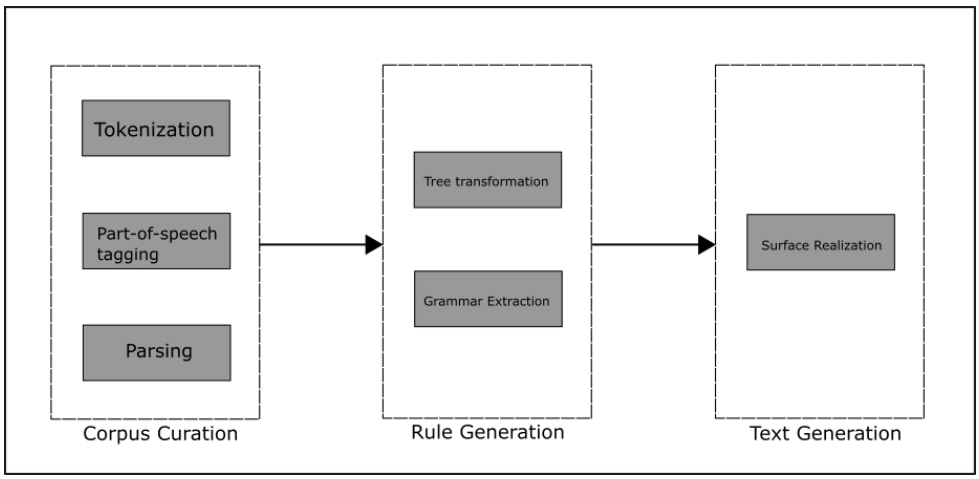


Figure 2: Overview of corpus-based NLG system

A major advantage of using PLTAG over other more popular grammars such as Context-Free Grammar (CFG) is that PLTAG has the property of being re-writable at non-terminals nodes [24]. This is particularly useful for augmenting -with more functional features- the grammars that are induced from corpora whose texts may not all be well-formed. Using Natural Language Took Kit (NLTK), the transcribed commentaries are first pre-processed by splitting into sentences, annotated with part-of-speech tags and then parsed. The trees from the parsed sentences are then used to obtain probabilistic lexicalized tree-adjoining grammar rules for text generation. The task of an NLG system is to map from some input data to output text. To complete the text generation

task, input data from in-game events during the moment of highlights are first extracted. The trees that cover the input data are combined using unification and substitution techniques. The production rules that derive the combined trees containing all or most part of the input data are used to generate the text. Finally, we used simpleNLG [25] an off the self NLG toolkit to create and enforce all linguistic rules for the final text that describes the moment of highlight.

4. CORPUS CREATION

In the absence of naturally occurring corpus of human-written match highlights, since casters and analysts normally run commentaries orally we collected our own corpus of human transcribed audio post-match analysis. Researcher in natural language such as [26] argued that in the absence of linguistic rules both template-based and rule-based generation system can profit from the use of statistical information derived from corpora. Corpus curated from commentaries preserve the lexical and syntactic properties of words and sentences used by the casters and are natural to read and understand in contrast to other sources such as event logs or other human-written templates. To curate high-quality text from commentaries and post-match analysis of highlights moments, we considered only matches from professional tournaments under the Major championships category. We provide the details of the number of comments made by analysts and casters covering the selected matches in table 1. We selected 55 different professional Dota 2 matches across 5 tournaments in 2019 and 2020. First, we randomly selected the publicly available Dota 2 match highlights and powerplays from internet-based video on demand systems. We note here that powersplay analysis are short moments when analysts give post-match analysis on few selected moments of highlights. Comments made in powerplays analysis are considerable fewer than the comments made by casters during the live match.

Table 1: Summary of Dota 2 matches selected for corpus creation

	Tournament	**Games**	**Teams**	**Comments**
Powerplays	ESL One Mumbai 2019	6	several	154
	ESL One Katowice 2019	15	several	369
	ESL One Birmingham 2019	5	several	117
	ESL One Hamburg 2019	8	several	171
Commentaries	ESL One LA 2020	3	Nigma vs VP Prodigy	596
	ESL One LA 2020	2	Ehome vs RNG	605
	ESL One LA 2020	3	Fnatic vs Adroit	661
	ESL One LA 2020	2	OG vs Secret	766
	ESL One LA 2020	2	OG vs Vikin	330
	ESL One LA 2020	2	OG vs VP Prodigy	485
	ESL One LA 2020	4	RNG vs Newbee	542
	ESL One LA 2020	3	Secret vs Alliance	482

We selected powerplay analysis for all the 2019 tournaments and commentaries for all the 2020 tournaments. Five transcribers were paid to transcribe audio files extracted from the highlight moments into text documents with timestamps. One researcher with Dota 2 expert knowledge further improved the transcription quality by filling in all missing terms in the transcriptions.

5. EXPERIMENT DESIGN

Recent studies using quantitative and qualitative data from experiments on data-driven experiences for esport have shown that users want a holistic cross-platform experience [6]. We contribute to this finding by investigating if spectators that use text recaps to enhance their viewing experience benefit more than users that use the generated text recap in-place of viewing the highlights. To start our investigation, we asked *if text generated from the grammar induced from commentary corpus can accurately recap highlight moments in Dota 2 matches*. As a follow up question to study the utility of the generated text, we asked *if the text recaps generated by the derived grammar support understanding of highlight moments in Dota 2 matches.*

For our experimental setup, we collected quantitative data from 127 participants to address our question about linguistic quality of text generated from a corpus of commentaries and how the text generated support understanding of highlight moments in dota 2 matches. To test for evidence of highlight understanding, we created three different groups from the participants with three viewing condition to simulate different experiences. In the following section, we detail the choices we made for our design and the general procedure for our experiment.

5.1. Highlight Moment Selection

Esport tournament organisers typically organize the online broadcast of games into two streams to cover all the games leading to the final game of the competition. In order to evaluate generated text for high-quality moments of highlight, we selected the first game of the final event of the most recent professional competition at the time of conduction this experiment. Dota 2 is a multiplayer online battle arena (MOBA) game and as many other games in this genre players strategically build their heroes through the cause of a game in order to increase their teams win probabilities. Generally, there are three loosely defined phases of the game and many casters and analyst talk about highlight moments in each game based on how far it has progressed. The highlights in the first (early) phase typically progress for about 15 minutes according to [27], and it includes events where players are gaining experiences (XP) and golds mostly from NPC to unlock levels for their heroes. During the second phase of the game, highlights shifts to groups of players as they begin to purchase items from gold earned and engage in fights as a team. Highlights in the third phase are more coordinated team fights with mostly all players fully contributing towards defeating their opponents. In selecting highlight moments for evaluating our NLG system, we focus on the second phase of the game where events that have to be described in the text are within reasonable size for participants judging the generated text. We selected the first two highlights starting at minute 15:45 and 17:07 after the first phase of the game for investigating our research question on the accuracy of generated text and how they support understanding of the highlight moment.

5.2. Evaluating Generated Text

To evaluate the accuracy of generated text and their potential to support understanding of highlight moments, we align our approach with current research such as [28] and [3] who emphasized using human evaluation over automatic metrics. Human evaluation using Likert scale has been shown to give a better evaluating of text accuracy of NLG systems designed to generate short texts. Furthermore, we ensured that the quality of evaluation we get from the participants is not reduced by restricting our crowdsourcing platform to Prolific. Prolific has been reported by [29] to have a pool of participants that is more diverse and honest than other crowdsourcing alternatives such as Amazon Turk. We recruited 127 participants from Prolific after sorting

responses according to the primary inclusion criteria of being active Dota 2 players and the ability to read and write in English language.

We divided our online survey into 2 parts, with each part corresponding to the two research questions we were investigating. In the first part, all 127 participants were first asked to watch a video clip of approximately two minutes with a Dota 2 highlight moment. After seeing the video clip, all the participants were shown the automatically generated text recap for the highlight moment in the video clip. Multiple-choice questions asking participants to rate the text on a 5-point Likert scale were then presented to participants to provide quantitative data on accuracy and fluency. For example, on the linguist quality of accuracy of the text recap we asked *"How accurately do you think the generated text recaps the highlight in the video that you watched previously?"*

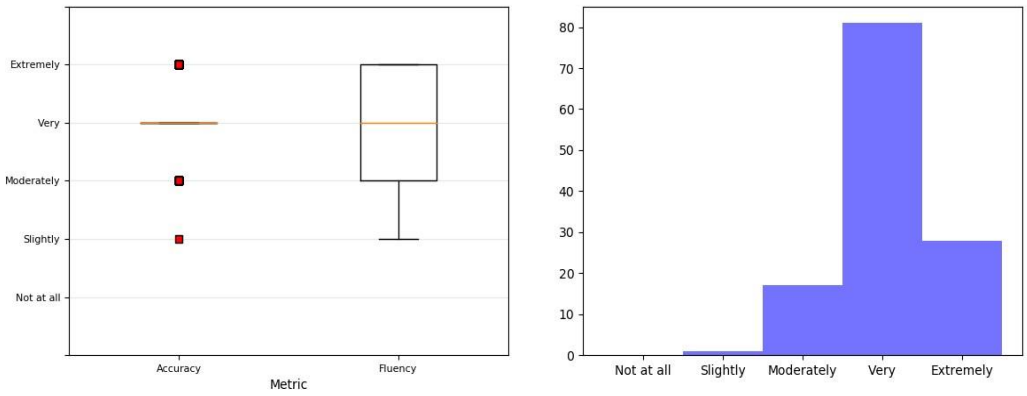
To evaluate the utility of the generated text in terms of support for the understanding of highlight, we divided the participants into three groups and presented each group with one of three conditions. The first group were shown a video clip only with a Dota 2 highlight moment and asked a set of multiple-choice questions about some of the events in the highlight. One of the questions for example was *"Who was the first hero to by killed in the highlight?"*. Results on data collected from participant response in this first group served as the baseline for assessing if there has been a gain in understanding of the highlight moment. The second group were shown the generated text only, while the third group were first shown the video clip followed by the generated text recapping the highlights. Both the second and third group were asked the presented with the same set of questions as the first group. In addition to the set of multiple-choice questions, a rank-based question was also asked of participants in all the three groups. We included a rank-based question based on the growing body of evidence such as in [30] and [28] that they provide a better way to access useful of texts generated from NLG systems.

6. RESULTS

In addressing our questions about automatically generated texts that recap esport highlights, our expectations from the results of our experiment were in two-fold. First, we expected our result on the linguistic quality of the generated text to align closely with outcomes of other work that reported high accuracy for templates and rule-based NLG approaches. Secondly, we expected that participants in our baseline group (Video only group) to perform better at recall tests than those in the group that received text only. Video recaps give extra cues such as visual and audio effects that may aid understanding. We expect that our result shows whether groups that read a text recap to enhance the viewing of video clip had a better performance on the recall test than the groups that either read the text recap only or viewed the video clip only.

6.1. Intrinsic Evaluation

In the first set of questions, our goal was to evaluate the NLG system using conventional intrinsic techniques where our human subjects read and rate the text generated after been shown a video highlight and its corresponding text recap. Debate continues about which property of text generated by an NLG system should be measured for linguistic quality. We followed best practices suggested by [28] for evaluating real-life of NLG system which placed text accuracy and fluency as top criteria for measuring the quality of the automatically generated text.



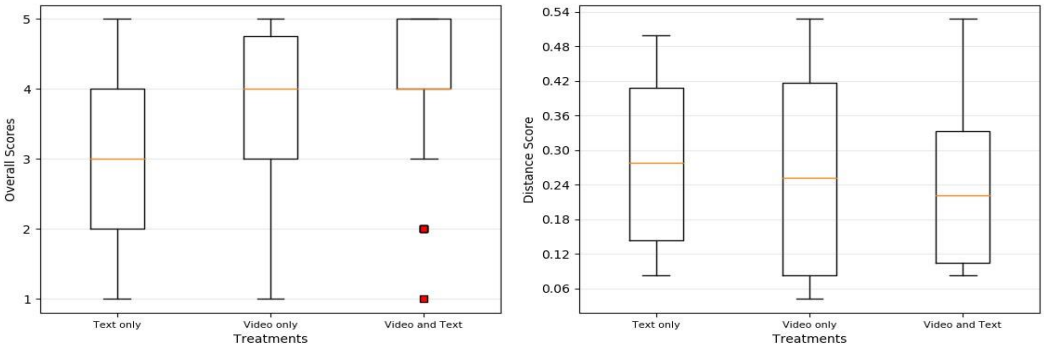
(a) Boxplot of responses to accuracy and fluency (b) Distribution of responses to questions on accuracy

Figure 3: Linguistic quality of generated text recaps by human evaluators

Looking at figure 3(a), it is clear from the small range in the boxplot for accuracy of generated text that the responses were more consistent and largely focused around the *"very accurate"* option. The data in figure 3(a) can be contrasted with data in chart of figure 3(b) where 64% of the 127 participants recruited selected the option of very accurate to indicate their perception of the accuracy of the generated text. The remaining 36% expressed their perception on the accuracy of text as *"extremely accurate 22%"*, *"moderately accurate 13%"* and *"slightly accurate"* at just about 1%. None of the participants selected the option of *"not accurate at all"*. On the question about fluency, the boxplot of figure 3(a) shows that the spread of participant response was not as consistent as responses for accuracy. The responses were skewed towards options that indicate the generated text have a higher degree of fluency.

6.2. Extrinsic Evaluation

In the second set of questions, we used extrinsic NLG evaluation techniques to measure the utility of the text generated in terms of how much they help support understanding of the match highlight event. Specifically, we presented the five task-based questions to the three groups of our human subjects and performed statistical analysis on the responses that we collected from the participants. First, we report the descriptive statistics of extrinsic methods in the boxplot of figure 4 and followed by statistical inferences using factorial design.



(a) Boxplot of recall test responses (b) Boxplot of recall test responses with ranking

Figure 4: Recall test responses for tasks and ranking question

The range in the scores in the boxplots of figure 4(a) shows different responses to five task presented to participants in the three different groups. The spread of responses from the group that read the text recap only shows an even distribution around a median score of 3. Responses from the groups that received video only and those that received video and text both had a median score of 4 and a spread of responses skewed towards higher scores. The range of responses for the group that received both video and text was smaller than those that received video only and therefore more consistent. A factorial experiment (One-way ANOVA) was conducted to validate the significance of the differences between the three groups. There was a significant difference at $p < 0.05$ in responses of the participants. Our experiment gave a F-statistic ($F_{2,123}$=3.535) and a p-value ($p = 0.0322$) for the three groups. A post hoc Tukey test showed that the score for the group that watched the video and read a text recap was significantly higher than the group of participants that read the text recap only. According to the results of group differences obtained from the Tukey test, there was no significant difference between the baseline group of participants that watched the video highlight only and the other two groups.
To further test if the understanding of participants who read a text recap only differed significantly from those who watched the video highlight and read the text recap, we presented a task-based test that required participants to rank their response. A distance score was calculated based on how different responses of the group participants was from the expected answer. Specifically, we used the Jaro–Winkler distance as metric for measuring the difference between participants' rank sequence and the correct rank sequence. The lower the score the more similar the ranking is between the response and the correct sequence. The range of distance score is shown in figure 4(b), while the boxplot showed some difference in the median of distance scores for participants in the three groups, a one-way ANOVA could not validate the difference shown in figure 4(b).

7. DISCUSSIONS

Our focus in investigating if grammar rules from casters' commentary corpus can create texts that accurately recap moments of highlight in Dota 2 matches was not to establish the general validity. We wanted to extend prior findings such as [23] and [24] on the general use of rule-based systems to domain-specific (esport) text generation task and explore the use of curated commentary corpus as a viable linguistic resource for such purpose. The high rating in the quantitative data of our study for accuracy and fluency from the first part of our experiment addresses the question about the linguistic quality of generated text recap from commentary corpus.

The result of our study did not find any evidence to validate our second hypothesis, where we asked if text-based recaps can supports understanding of highlight moment in Dota 2 match. The lower performance of participants who received text only in the recall tests is interesting, but not surprising. While the performance of the *"text-only"* group was less than the groups that viewed video highlights, we recognise a value for text-only recaps in the same way large-scale esport project like [6] identified that there are some circumstances where people are unable to view video and therefore may benefit from text recaps (e.g. during transit or when working and trying to keep up with a tournament). For these cases, the text recaps which participants perceived to be accurate in the first part of the study offer good value to esport spectators. In the following section, we discuss the findings of our study on NLG systems that generate text recaps with respect to their implications for real-life applications in esport.

7.1. Text Recap and Highlight Accuracy

One prospective benefit of generating text-based recaps using an NLG system is that they make match data more explicable and therefore more promotable. This means that text-based recaps can contribute to the democratisation of the performance data that esport content producers want to communicate to all their audiences regardless of skill level. According to [31], in many commercial NLG applications, it is essential that the system produce reasonable texts in all scenarios. A system that generates a text summary of a player's performance/skill from non-linguistic data should not produce incorrect text that may negatively impact their skill level. Building applied NLG systems should involve testing processes that evaluate the systems for performances in both expected outcomes and edge-case scenarios. In addressing the first question about linguistic quality of the generated, the high concentration of participant choice (64%) around the *"very accurate"* option on our 5-point Likert scales aligns with results in rule-based NLG literature such as [32] and [33] where templates and hand-engineered systems generally resulted in semantically accurate text than systems that use machine learning approaches like neural text generation. The finding on the accuracy of the generated text recaps contrast with our finding on participant response for the fluency question, where the largest response was also for *"very fluent"* but only by 38% of the participants. This is however consistent with recent large-scale NLG system evaluation by [34], involving 20 different NLG systems. The evaluation of the shared NLG task found that systems that produce semantically accurate text often suffer from lack of naturalness (fluency and readability) in the text if controls are not applied before the generation process. Investigation of NLG metric by [32] suggested that correct semantic accuracy is regarded by users as more important than fluency and should be prioritised when training the models.

Accuracy of text is particularly important in building applied systems for domains where texts are generated to reduce complexity and to provide simple and clear information to audiences. Consider esport where a large-scale case study of a production tool called Echo was deployed. Echo uses live and historic match data to detect extraordinary player performances in the popular esport Dota2 and dynamically translates interesting data points into audience-facing graphics superimposed with text. Echo was deployed at one of the largest yearly Dota 2 tournaments, which was watched by 25 million people. The creator of Echo [8] analysed 40 hours of video, 46,000 live chat messages, and feedback of 98 audience members and showed that Echo measurably affected the range and quality of storytelling, increased audience engagement, and invoked rich emotional response among viewers. In one instance [8] reported that a small number of twitch users picked part of the text in the graphic and expressed concern with a phrase that casters had relayed as a global record instead of a personal record that referenced a single player's history. [8] reported that there were only three total instances among 5091 chat posts and their evidence suggests that a small number of users do inspect little details in the small texts, the query of the validity of the details. According to [8], they included records of matches after a major patch (update that changes the game) because the patch significantly impacted many performance features that were tracked by Echo. This made it questionable to compare data collected before and after the patch.

Patch updates and how they impact on the performance of players is often an issue that is discussed by casters during the game and by analysts during the post-match analysis. Extracting grammar rules from corpus curated from transcripts of previous caster and analyst commentaries in a similar way to our methodology can help produce a textual explanation of player performance that account for the type of edge case that appeared in Echo's graphics. Esport data analytic tools such as Echo can be developed into second-screen companion apps with a functional NLG module that handles accurate processing of historic and real-time non-linguistic performance data and generation of textual messages for esport audience.

7.2. Text Recap and Highlight Understanding

According to [2], the video games being played in eSports e.g. Dota 2, are usually complex and require a considerable amount of concentration to comprehensively follow the game. While the simplicity that comes with the descriptions of the text may bring to focus the outcomes of key events during a highlight moment, they lack the visual and audio signals that help the audience understand more of what is happening. As reflected in the better recall test scores of participants that saw both video highlight and text recap, a case to improve audience understanding of Dota 2 can be made to take advantage of both the simplicity of the generated text and additional cues from video by bundling them together when recapping highlight moment. One possible application scenario of this finding is the addition of an explainer feature to already existing systems such as the WEAVR app in [6]. The WEAVR app annotates Dota 2 maps with simple explanations of the highlights and stats from in-game data. Enhancing these annotations with engaging text that explains the context of these highlights to new players of the game makes the application of our findings in this part of our study an interesting one.

The continuous improvements in both hardware and middleware used for delivering online broadcasting remain the driving force for esports and video game streaming as a new form of online media consumption. Acquisition of knowledge related to an esport game has been shown by [2] to contribute the most to the frequency of watching esports. The audience of esport who are presented with simple information about intense moments in a game in a way that supports acquiring knowledge such as text recaps of highlight moments are therefore less likely to be part of the churn that typically occur at the onset of mastering a game. In the face of competition for viewership, the esport content producers trying to stand out and attract audiences to their platform can take advantage of the knowledge acquisition capability that generated text recaps can enable when superimposed on a video highlight. An NLG system that is well integrated and adequately recap highlights of visually complex moments in esport games can help increase audience engagement and loyalty.

In addition to the growing online viewership, the current decade-leap projected by [35] and [36] in the consumption of digital media due to the global pandemic may provide the spotlight that esports needs to encourage broader participation from mainstream television broadcasting. One of the ways the result of this study will benefit adoption by mainstream TV is the ease of integration with the way mainstream broadcasting deliver headlines and breaking news. Many major Mainstream TV channels have decades of experience at relaying details of breaking new by presenting the information as new tickers scrolling across the area at the bottom of the screen. Automatically generated text recapping complex in-game events can be scrolled as breaking stories for the time the video recap is being broadcast on the screen.

7.3. Limitation

As an alternative approach to a neural text generation, our grammar-based approach has avoided the need for large labelled data required by neural network to generate accurate text. However, the process of obtaining the generation grammar from the commentary corpus meant that all audio recordings of the commentaries had to be transcribed first. Gameplay in Dota 2 is extremely complex [8] and as reported by [37], Dota 2 lies on the threshold between excitement and difficulty. Dota 2 require a lot of domain knowledge and expertise beyond the skill of a proficient transcriber. As result, the process of curating a commentary corpus was a resource-intensive task requiring a two-tiered (non-expert and expert) approach to improving the transcribed commentaries. It will be interesting to see if this task can be less intensive when state-of-the-art speech-to-text machine learning models such as reported in [38] are used for transcribing commentaries.

8. CONCLUSIONS

In this paper, we have induced rules from a collection of annotated esport commentary for the task of generating accurate text that recaps highlight moments in Dota 2 matches. The experiments in our study have shown that viewers perceive rule-based text recaps as accurate and fluent accounts of highlight moments in Dota 2 matches. Comparisons between three groups of participants were made to test the utility of text-based recaps implemented using an NLG system. Using recall tests, we observed significant improvements in the score of participants that viewed video highlights and its text recap over the groups that viewed the text recap only. Consequently, we can conclude from our results that esports commentaries are a viable linguistic resource for generating accurate text recaps that can support understanding of the complex actions in Dota 2 video highlights. Further to recapping video highlights, the NLG system used for the text-based recaps can also add to spectator experiences by making lexical choices according to the skill and preference of the viewer. We intend to explore this variation of the NLG system for creating personalised text recaps in future work.

ACKNOWLEDGEMENTS

This work has been created as part of the Weavr project (weavr.tv) and was funded within the Audience of the Future programme by UK Research and Innovation through the Industrial Strategy Challenge Fund (grant no.104775) and supported by the Digital Creativity Labs (digitalcreativity.ac.uk), a jointly funded project by EPSRC/AHRC/Innovate UK under grant no. EP/M023265/1.

REFERENCES

[1] Lucas Kempe-Cook, Stephen Tsung-Han Sher, and Norman Makoto Su, (2019) "Behind the voices: The practice and challenges of esports casters". In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019*, pp565.

[2] Juho Hamari and Max Sjoblom, (2017) "What is esports and why do people watch it?", *Internet Research*, Vol. 27, No. 2, pp211-232.

[3] Craig Thomson and Ehud Reiter, (2020) "A gold standard methodology for evaluating accuracy in data-to-text systems", *ArXiv preprint arXiv:2011.03992.*

[4] Jason G. Reitman, Maria J. Anderson-Coto, Minerva Wu, Je Seok Lee, and Constance Steinkuehler, (2020) "Esports research: A literature review", *Games and Culture*, Vol. 15, No. 1, pp32–50.

[5] Douglas Heaven, (2014) "Rise and rise of esports", *New Scientist*, Vol. 223, No. 2982, pp17.

[6] Athanasios Vasileios Kokkinakis, Simon Demediuk, Isabelle Nolle, Oluseyi Olarewaju, Sagarika Patra, Justus Robertson, Peter York, Alan Pedrassoli Pedrassoli Chitayat, Alistair Coates, Daniel Slawson, Peter Hughes, Nicolas Hardie, Ben Kirman, Jonathan Hook, Anders Drachen, Marian F Ursu, and Florian Block, (2020) "Dax: Data-driven audience experiences in esports". In *ACM International Conference on Interactive Media Experiences*, IMX '20, pp94–105.

[7] Thomas Horky and Philipp Pelka, (2017) "Data visualisation in sports journalism", *Digital Journalism*, Vol. 5, No. 5, pp587–606.

[8] Florian Block, Victoria Hodge, Stephen Hobson, Nick Sephton, Sam Devlin, Marian F. Ursu, Anders Drachen, and Peter I. Cowling, (2018) "Narrative bytes: Data-driven content production in esports". In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video, TVX '18*, pp29–41.

[9] Christoph Eggert, Marc Herrlich, Jan Smeddinck, and Rainer Malaka, (2015) "Classification of player roles in the team-based multi-player game dota 2". In Konstantinos Chorianopoulos, Monica Divitini, Jannicke Baalsrud Hauge, Letizia Jaccheri, and Rainer Malaka, editors, *Entertainment Computing - ICEC 2015*, pp112–125.

[10] Simon Demediuk, Peter York, Anders Drachen, James Alfred Walker, and Florian Block, (2019) "Role identification for accurate analysis in dota 2". In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 15, No. 1, pp130–138.

[11] K. Conley and D. Perry, (2013) "How Does He Saw Me? A Recommendation Engine for Picking Heroes in Dota 2", Available: http://cs229.stanford.edu/proj2013/PerryConley-HowDoesHeSawMeARecommendationEngineForPickingHeroesInDota2.pdf, [Accessed: 10-Oct-2020].

[12] Tobias Mahlmann, Matthias Schubert, and Anders Drachen, (2016) "Esports analytics through encounter detection", In *Proceedings of the MIT Sloan Sports Analytics Conference*. *MIT Sloan*.

[13] Wei-Ta Chu and Yung-Chieh Chou, (2015) "Event detection and highlight detection of broadcasted game videos". In *Proceedings of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, HCMC '15, pp1–8.

[14] Yale Song, (2016) "Real-time video highlights for yahoo esports", *ArXiv*, Vol. abs/1611.08780.

[15] Charles Ringer and Mihalis A. Nicolaou, (2018) "Deep unsupervised multi-view detection of video game stream highlights". In *Proceedings of the 13th International Conference on the Foundations of Digital Games, FDG '18*.

[16] Justus Robertson and Michael Young, (2014) "Gameplay as on-line mediation search", *Experimental AI in Games*.

[17] Shaowen Bardzell, Jeffrey Bardzell, Tyler Pace, and Kayce Reed, (2008) "Blissfully productive: Grouping and cooperation in world of warcraft instance runs". In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pp357–360.

[18] Yubo Kou and Xinning Gui, (2014) "Playing with strangers: Understanding temporary teams in league of legends", *CHI PLAY*, In *Proceedings of the 2014 Annual Symposium on Computer-Human Interaction in Play*, pp161–169.

[19] Marco Furini and Manuela Montangero, (2016) "Tsentiment: On gamifying twitter sentiment analysis". In *2016 IEEE Symposium on Computers and Communication (ISCC)*, pp91–96.

[20] Yoonsung Hong, Haewoon Kwak, Youngmin Baek, and Sue Moon, (2013) "Tower of babel: A crowdsourcing game building sentiment lexicons for resource-scarce languages", In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pp549–556.

[21] Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang, (2020) "Few-shot nlg with pre-trained language model", *CoRR*, Vol. abs/1904.09521.

[22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, SandhiniAgarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, (2020) "Language models are few-shot learners", *ArXiv*, Vol. abs/2005.14165.

[23] Huayan Zhong and Amanda Stent, (2005) "Building surface realizers automatically from corpora using general-purpose tools". In *Proceedings of UCNLG'05*, pp49–54.

[24] Trevor Cohn, Sharon Goldwater, and Phil Blunsom, (2009) "Inducing compact but accurate tree substitution grammars". In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp548–556.

[25] Albert Gatt and Ehud Reiter, (2009) "Simplenlg: A realisation engine for practical applications", In *Proceedings of the 12th European Workshop on Natural Language Generation, ENLG '09*, pp90–93.

[26] Kees Van Deemter, Emiel Krahmer, and Mariet Theune, (2005) "Real versus template-based natural language generation: A false opposition?" *Comput. Linguist.*, Vol. 31, No. 1, pp15–24.

[27] Kevin Godec, (2011) "Dota 2 guide: Welcome to dota, you suck", *Purge gamers*, Available: https://purgegamers.true.io/g/dota-2-guide, [Accessed: 05-Nov-2020].

[28] Chris Van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer, (2019) "Best practices for the human evaluation of automatically generated text". In *Proceedings of the 12th International Conference on Natural Language Generation*, pp355–368.

[29] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti, (2017) "Beyond the turk: Alternative platforms for crowdsourcing behavioral research", *Journal of Experimental Social Psychology*, Vol. 70, pp153 – 163.

[30] Jekaterina Novikova, Ondˇrej Dusek, and Verena Rieser, (2018) "RankME: Reliable human ratingsˇ for natural language generation". In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2, pp72–78.

[31] Ehud Reiter and Robert Dale (2002). Building applied natural language generation systems. *Natural Language Engineering*, Vol. 3, No. 1, pp57-87.
[32] Ehud Reiter and Anja Belz, (2009) "An investigation into the validity of some metrics for automatically evaluating natural language generation systems", *Computational Linguistics*, Vol 35, No. 4, pp529–558.
[33] Sam Wiseman, Stuart Shieber, and Alexander Rush, (2017) "Challenges in data-to-document generation". In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp2253–2263.
[34] Ondˇrej Dusek, Jekaterina Novikova, and Verena Rieser, (2020) "Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge", *Computer Speech & Language*, Vol. 59, pp123 – 156.
[35] Bidit L. Dey, Wafi Al-Karaghouli, and Syed Sardar Muhammad, (2020) "Adoption, adaptation, use and impact of information systems during pandemic time and beyond: Research and managerial implications", *Information Systems Management*, Vol. 37, No. 4, pp298–302.
[36] Netta Iivari, Sumita Sharma, and Leena Venta-Olkkonen, (2020) "Digital transformation of everyday life – how covid-19 pandemic transformed the basic education of the young generation and why information management research should care?", *International Journal of Information Management*, Vol. 55, pp102183.
[37] Nick Yee, (2016) "Game Genre Map: The Cognitive Threshold in Strategy Games", *Quantic Foundry*, Available: https://quanticfoundry.com/2016/01/20/game-genre-map-the-cognitive-threshold-in-strategy-games/, [Accessed: 05-Oct-2020].
[38] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed and Michael Auli, (2020) "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *ArXiv preprint arXiv:2006.11477*.

AUTHORS

Oluseyi Olarewaju is a Postdoctoral Research Associate in the Department of Theatre, Film, Television and Interactive Media at the University of York. UK. He is a member of Weavr within Digital Creativity Labs. Oluseyi completed his PhD in Computer Science at University of Brighton focusing on Personalization Systems. His research interests are in Cross-domain Recommender Systems, Predictive Modelling and Natural Language Processing. Oluseyi is currently exploring the application of computational linguistics for generating personalized stories from game.

Jon Hook is a Lecturer in Interactive Media at the University of York's Theatre, Film, TV and Interactive Media department. He has a background in computing, having done a PhD in Human-Computer Interaction at Newcastle University. Jon's research explores the design and development of novel interactive technologies for abroad range of artistic and everyday creative practices. In the past he's investigated the design of technologies that help live performers, digital artists, design educators and people with disabilities be creative in new ways. In DC Labs he's continuing this work through collaborations with high-profile partners in interactive television, theatre performance and digital heritage.