There's only one way to find out.

– Greg Lee.

University of Alberta

AUTOMATED STORY-BASED COMMENTARY FOR SPORTS

by

Gregory Michael Kenney Lee

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

Fall 2012
Edmonton, Alberta

Abstract

Automated sports commentary is a form of automated narrative and human-computer interaction. Sports commentary exists to keep the viewer informed and entertained. One way to entertain the viewer is by telling brief stories relevant to the game in progress. We introduce a system called the Sports Commentary Recommendation System (SCoReS) that can automatically suggest stories for commentators to tell during games. Through several user studies, we compared commentary using SCoReS to three other types of commentary and showed that SCoReS adds significantly to the broadcast across several enjoyment metrics. We also collected interview data from professional sports commentators who positively evaluated a demonstration of the system. We conclude that SCoReS can be a useful broadcast tool, effective at selecting stories that add to the enjoyment and watchability of sports. SCoReS is a step toward automating sports commentary, and thus automating narrative.

Acknowledgements

I would like to thank my supervisor, Vadim Bulitko, for his support and guidance throughout my graduate career. In particular, I would like to thank him for helping me switch Ph.D. topics to something I am passionate about, making my studies more enjoyable and fulfilling.

Thanks to my wife, Melissa for her love and encouragement at all times, but in particular during the last few (very busy) months leading to the submission of this document. Also, for taking on all the "parental responsibilities" during the writing of this document. Thanks to my son, Henry, for never ceasing to ask me to 'come ow-side' even after he noted that 'daddy work, daddy work'. Thanks to my daughter, Charlotte, for smiling at the right times.

Thanks to Elliot Ludvig for his many helpful suggestions concerning both the Psychology side of the experiments, and the baseball side. Also thanks for helping guide the direction of the project to something meaningful for AI.

Thanks to Dr. Brenda Trofanenko for her tremendous support of my research after my move to Nova Scotia, including the use of her office, meeting room and her constant reminders that "The end is in sight!".

Thanks to the faculty and staff at both the University of Alberta and Acadia University for providing me with all the resources I needed for my studies. In particular, thanks to the Computing Science and Psychology departments at both institutions.

Thanks to the IRCL research group for your help throughout the years. Thanks especially to David Thue - the MVP of the group. Thank you to NSERC, iCORE, Alberta Ingenuity, the Government of Alberta and the Department of Computing Science for funding me throughout my time at the University of Alberta.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Sports broadcasting is a billion-dollar industry. Most professional sports are broadcast to the public on television, reaching millions of homes. The television experience differs in many ways from the live viewing experience, most significantly through its commentary.

Much research has been done into the importance of commentary during sports broadcasting. When watching a game on television, "...the words of the commentator are often given most attention" [Duncan and Hasbrook, 1988]. The commentary has the effect of drawing the attention of the viewer to the parts of the picture that merit closer attention [Kennedy and Hills, 2009], an effect called *italicizing* [Altman, 1986]. Commentary can also set a mood during a broadcast. A commentator who creates a hostile atmosphere during a broadcast often makes the viewing experience more enjoyable for the viewer [Bryant *et al.*, 1982]. The descriptions given in a broadcast are so useful that fans often bring radios to live games in order to listen to the interpretations of the commentators [Michener, 1976]. Also, some sporting venues now support a handheld video device that provides the spectator with in-game commentary [Ross, 2012].

The purpose of commentators is to help the viewer follow the game and to add to its entertainment value. One way to add entertainment to a broadcast is to tell interesting, relevant stories from the sport's past. The sport of baseball is particularly suited to storytelling. Baseball is one of the oldest professional sports in North America, existing since 1876. This longevity provides a rich history from which to draw interesting stories. Dick Enberg, a sports commentator of various sports for over 50 years once claimed of baseball, "It's a great broadcaster sport. It's the best broadcaster sport!". The more popular commentators are known as "storytellers" [Smith, 1995], as they augment the games they call by adding stories that connect baseball's past to its present. One of these "storytellers", Vin Scully, has been commentating for Brooklyn/Los Angeles Dodgers games since 1950 and has been

voted the "most memorable personality in the history of the franchise".

A typical baseball game lasts about three hours, but contains only ten minutes of action, where the ball is live and something is happening on the playing field. This leaves two hours and 50 minutes of time where little is happening on the playing field, and it is the job of the commentators to entertain the viewer. This downtime is a good time to tell stories [Ryan, 1993]. Baseball is also known for being a statistically dense league, and being able to match statistics from the current game state to a past situation in baseball adds to the commentators ability to entertain the viewer.

To illustrate, consider the case where one baseball team is trailing by four runs in the ninth inning. As this is not a particularly interesting situation, it may be a good time for a story. An appropriate story might be that of the Los Angeles Dodgers, who on September 18, 2006, were also trailing by four runs in the bottom of the ninth inning. The Dodgers hit four consecutive home runs to tie the game. The broadcast team could tell this story to the viewer, because the situations are similar. Thus, what is needed is a *mapping* from a game state (a particular point in a game) to an appropriate story.

Sports storytelling is a form of narrative discourse. Narrative discourse is a creative activity that involves relaying to an audience a series of events in an interesting and entertaining manner. It is a recounting of contingent events from the past with one or more main characters. Automating narrative discourse is a challenging problem for Artificial Intelligence (AI) and a subject of much recent research [Young, 2007]. This thesis explores the following hypothesis:

> Sports colour commentary via storytelling can be automated using Artificial Intelligence and Information Retrieval methods.

Specifically, we set out to test whether an AI approach can be developed that maps game states to relevant stories, thereby significantly increasing audiences' enjoyment of the broadcast. To test this hypothesis, we develop an AI system that tells stories in the context of baseball. To do so, the Sports Commentary Recommendation System (SCoReS) learns offline to connect sports stories to game states, provided with some scored examples of story-game state pairs. This learned mapping is then used during baseball games to suggest relevant stories to a (human) broadcast team or, in the absence of a broadcast team (e.g., in a sports video game), to autonomously output a relevant story to the audience. In the case of suggesting stories to commentators, SCoReS is an example of human-computer interaction [Dix *et al.*, 1993], as it is a computer system designed to communicate to humans

information that will improve their performance.

This thesis makes the following contributions to the field of AI. First, we formalize story-based sports commentary as a mathematical problem. Second, we use machine-learning methods and information retrieval techniques to solve the problem. This solution for the specific domain of sports story selection, but is also general enough to be used in other domains involving story selection. Third, we implement the approach in the domain of baseball, evaluate the resulting AI system by observing feedback from human participants and show that it is effective in performing two separate tasks: i) automating sports commentary, and thus automating narrative in a special case, and ii) assisting human commentators. That is, we show that our combination of information retrieval techniques is able to map previously unseen baseball game states to stories in a sufficiently effective manner to improve the enjoyability of baseball broadcasts, increase interest in watching baseball and suggest stories to professional commentators that they would tell during a live game.

The rest of the thesis is organized as follows. We next describe colour commentary in detail, and formulate the problem of mapping sports game states to interesting stories. This is followed by a review of research related to story-based commentary. Next, we describe information retrieval techniques in detail and describe our approach to mapping game states to stories – a combination of information retrieval techniques designed to rank stories based on a given game state, and then ensure the higher-ranked stories are indeed relevant to said game state. Following this, we describe empirical work performed to choose a story ranker, and then use this ranker to select stories for baseball broadcasts. The quality of the story mapping is evaluated with user studies and demonstrations to professional sports commentators. We conclude with a discussion of lessons learned, future research and applications, and a summary of the contributions of this thesis.

Chapter 2

Problem Formulation

Commentating in sports generally involves two people – a play-by-play commentator and a colour commentator. Play-by-play commentating involves relaying to the audience what is actually happening in the field of play. Beyond reporting the actions of the players as they happen, the play-by-play commentator typically mentions such facts as the score of the game, upcoming batters and statistics for the teams and players involved in the game. Colour commentary, on the other hand, is much more subjective and broad, with the purpose being to add entertainment (i.e, "colour") to the broadcast. This can be done in several ways, as we describe below.

After a play is over, colour commentators tend to analyze what has happened beyond the surface. For instance, if a player swings awkwardly at a pitch and misses, the colour commentator may point out that the reason for the hitch in his swing is that he has an ankle injury, and is unable to plant his foot when swinging. This gives the viewer some extra information beyond what he or she can see, or is told by the play-by-play commentator.

Another manner in which the colour commentator adds to a broadcast is by giving background information on the players involved in the game. While the play-by-play commentator can provide a player's statistics, the colour commentator tends to add more personal information about a player, including possible interactions with that player.

Passing on their expertise in the sport at hand is one more way colour commentators add to the broadcast. An example of the expertise of a colour commentator could be seen in the July 23rd, 2011 game between the New York Yankees and Oakland Athletics. Paul O'Neill, a former Major League Baseball (MLB) player, who played for the Yankees among other teams, was the colour commentator for this game for the Yankees Entertainment and Sports (YES) network. From the broadcast booth, he was able to pick up on a pattern of pitches from the pitcher for the Athletics, Rich Harden. He noted that Harden almost never threw

two fastballs in a row (an oddity for a major league pitcher), but that he would often throw consecutive change-ups, a deceptively slow pitch that appears to the batter as a fastball, often causing him to swing too early. After O'Neill pointed out this pattern, it was evident to the viewer that he was correct. As pitch speeds are displayed on YES, one can verify fairly quickly whether a pitch was a fastball or not.

Another rich aspect of colour commentary is storytelling, which is our focus. Effective storytelling in sports broadcasts involves telling a story that is interesting to the audience, and that is related to what is actually happening in the game being broadcast. While play-by-play commentators are generally trained journalists, colour commentators are typically former professional athletes or coaches. As former members of the league being shown, they are thought to bring a level of expertise to the broadcast. Because they actually played or coached in the league, colour commentators tend to tell stories from their own experiences in the game. This gives the audience a first-hand account of snippets of baseball history, which can be quite entertaining. Unfortunately, colour commentators do not have first-hand knowledge of most of baseball history. They can draw upon their background knowledge of the game to tell more stories, but they are limited in their memory as are all humans. Their knowledge of stories, however vast for human beings, is still limited relative to the total set of baseball stories available. This is where AI can be of assistance. Computers can both store many more stories than a human brain as well as quickly compute the quality of a match between each story in the library and the game state at hand.

The problem we are attacking in this thesis is to tell interesting stories as live commentary to a sports game. To compare stories to game states, we extract features from both, such as the score, the teams involved and what type of action is happening (i.e., a home run). Formally, the game state is a vector of n numeric features: $\vec{g} = (g_1, g_2, \ldots g_n)$. To illustrate: binary feature g_1 may be 1 if in game state $\vec{g}$ there is a runner on first base. Integer feature g_2 can be the current inning. Similarly, a baseball story can be described with a vector of p numeric features: $\vec{s} = (s_1, s_2, \ldots s_p)$. Binary feature s_1 can be 1 if the story involves a runner on first base and integer feature s_2 can be the inning number mentioned in the story. The task is then to map $\vec{g}$ to a relevant and interesting $\vec{s}$. Figure 2.1 demonstrates the problem. The game state is extracted from what is happening on the field and made available to the AI system. The story database is also provided to the system. Based on the game state, the AI system selects a relevant, interesting story, and relates it to the viewer.

We next review related work in automated commentary and storytelling, before using the above-described framework to tackle the problem of mapping game states to stories.

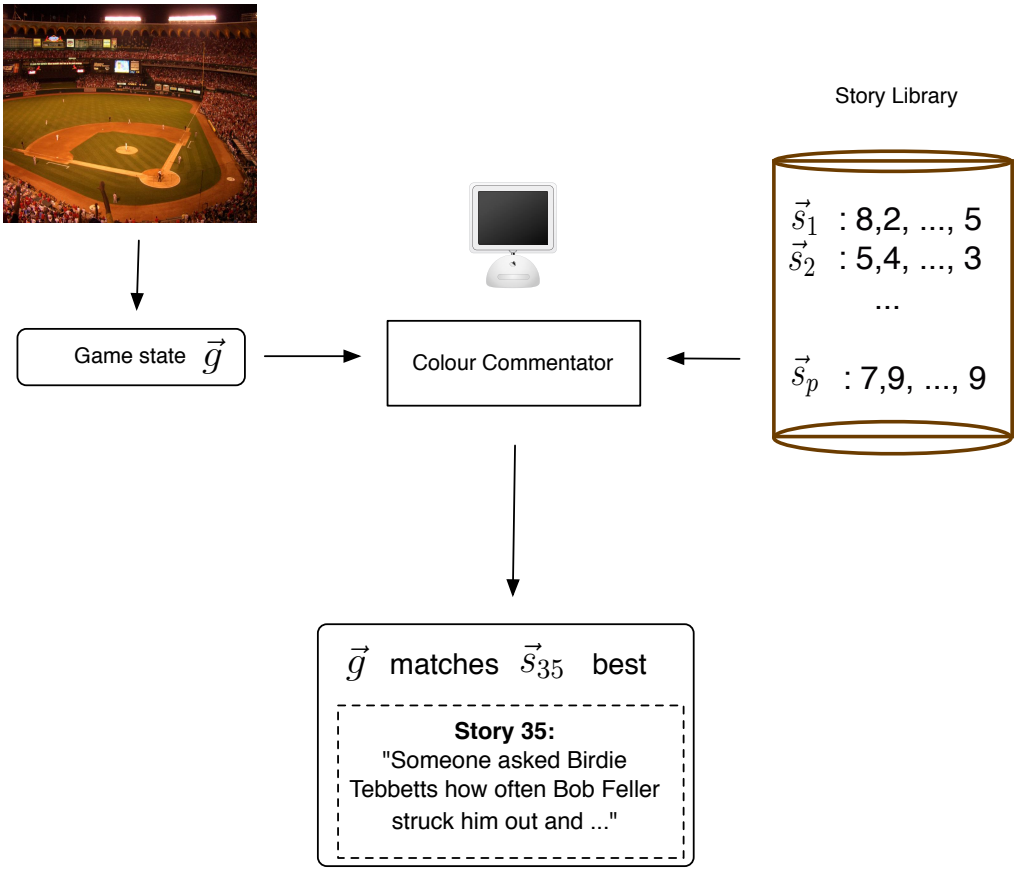

Figure 2.1: An AI colour commentator selects interesting stories for the viewer.

Chapter 3

Related Work

In this chapter, we review existing research relevant to the problem we are solving. Related work in automated and human storytelling is presented in Section 3.1, while automated play-by-play commentary, with some added colour commentary is presented in Section 3.2. None of the existing work delivers story-based colour commentary to a live sports game. Following this, we review research related to our solution to the problem, in the field of information retrieval (Section 3.3).

3.1 Automated and Human Storytelling

Narrative generation explicitly attempts to automate storytelling by having computers create a story with entertaining plot and characters. The ability to generate effective narrative is important in different applications, such as entertainment, training and education [Riedl, 2004], and this area has been studied for over twenty years [Hammond *et al.*, 1990]. Story generation is generally divided into three categories – character-centric (where simulated characters determine the story), author-centric (where an author's thought process is modeled) and story-centric (where the structural properties of stories themselves are modeled) [Bailey, 2009]. As narratives are a sequence of events describing how a story world changes over time, some techniques model story generation as a planning task [Riedl, 2004]. More recently, Analogy-based Story Generation has received attention. This approach generates new stories from existing stories. SAM (Story Analogies through Mapping) [Ontanon and Zhu, 2011] completes a partial story by transferring knowledge from a source story. SAM breaks stories into phases and generates an injective mapping from the phases of the source story to the phases of the partial story. SAM then finds analogous mappings between the two stories, based on similarities between the computer understandable descriptions of the source story and the target story. In sports commentating, the stories told

are true. Narrative generation approaches create fictional stories and are thus inapplicable to our problem.

There has been much work on automated storytelling in non-sports video games [Roberts and Isbell, 2008]. Some systems generate or adapt a story due to actions taken by the player, creating a so-called interactive drama. PaSSAGE (Player-Specific Stories via Automatically Generated Events) [Thue *et al.*, 2007] is a system that models the current user and adapts the story within the game to suit the system's model of the user. Values for five character attributes are assigned to the player, and then updated on the basis of the player's actions within the game. Automated Story Director (ASD) [Riedl *et al.*, 2008] is an experience manager that accepts two inputs: an exemplar narrative that encodes all the desired experiences for the player and domain theory encoding the rules of the environment. As the player can influence the narrative, ASD adapts the examplar narrative to ensure coherence of the story and a feeling of agency for the player. These adaptations are pre-planned based on the domain theory. Systems such as PaSSAGE and ASD actively change the course of the game based on user actions. As a commentary system cannot alter what happens in a live sports game (real or video game), these systems are not directly applicable to our problem.

Human commentators try to weave a narrative with a coherent plot through unpredictable sports games [Ryan, 1993]. Plot is a difficult narrative dimension for the broadcast team, as they do not know what the end result of the game will be until it concludes. They can, however, attempt to predict an ongoing theme based on what has happened so far, and begin to build up a plot that fits with the predicted ending. This brings into play different themes that the broadcaster can choose to follow, and he or she can follow more than one at a time (hedging his or her bets, so to speak). In the broad scheme of narratives, the range of themes spans all of human experience. In baseball, these themes are much more limited, and they include: the Incredible Come-From-Behind Victory, the Fatal Error, the Heroic Feat, the Lucky Break Victory, the Unlikely Hero, the Inevitable Collapse, Overcoming Bad Luck, Persistence That Pays Off, Last Chance, Futility, Wasted Opportunity and Opportunism [Ryan, 1993].

In similar work, Bryant [Bryant *et al.*, 1977] identifies 15 themes, which he labels as *motifs*: Spirit, Competition, Human Interest, Urgency, Pity, Miracle, Gamesmanship, Comparison, Performance Competence, Physical Competence, Old-College-Try, External Forces, History, Personnel and Glory. The work by both Ryan and Bryant involves human storytelling, not automated storytelling, so it is not applicable to our problem of automated

commentary. Also, the themes they present require more information than can easily be expressed by the set of numerical features used in our approach.

3.2 Automated Commentating in Sports

StatSheet [Allen, 2012] and *Narrative Science* [Frankel, 2012] automatically write previews for sports games that have not yet happened, and summaries about sports games that have already happened. For summaries, they are provided with statistics from a completed game and compose a narrative about the game, with the goal being to provide an interesting summary of game events. For previews, they are provided with statistics from past games, and compose a preview for the game that should entice the reader to watch said game. Neither Statsheet nor Narrative Science operate with live game data and both are unable to solve our problem of providing live stories during a game. They may, however, provide an additional potential database of stories about past games to augment our current system.

Modern commercial sports video games typically employ professional broadcast teams from television to provide commentary for the action in the game. This involves pre-recording large amounts of voice data that will be reusable for many different games. As an example, a clip to the effect of "Now that was a big out!" could be recorded and used in multiple situations where one team was in dire need of retiring a batter or runner. Recorded clips often use pronouns (i.e, "he" and "they") so that they are not specific to any particular player or team. This makes the commentary generic, which reduces the amount of recording required, as recording voice data can be time consuming and expensive. Unfortunately, generic commentary is less exciting to the audience. We would like our AI system to deliver colourful commentary tailored to the current game by mapping the current game state to a story chosen specifically for said game state.

The *MLB: The Show* [SCE San Diego Studio, 2009] suite of games is often considered to be at the leading edge of baseball video games. Recorded clips of Matt Vasgersian of the MLB network act as the play-by-play commentary while Dave Campbell of ESPN's Baseball Tonight is the colour commentator. Most of the colour commentary involves analysis of what has happened on the field and indirect suggestions to the player as to how to improve their play. As far as we have seen, there is no storytelling in these games, which is what our system is designed to add to sports commentary.

In *MLB 2K7*, another leading baseball video game, Jon Miller and Joe Morgan of *ESPN: Sunday Night Baseball* provide the play-by-play and colour commentary, respec-

tively. There is at least one instance of relating a game state to a story from baseball's past, occurring when a third strike is dropped by the catcher and he has to throw to first base to record the out. Here, Miller reminds us of a 1941 World Series game where Mickey Owens committed an error on a similar play, changing the course of the series. Figure 3.1 shows a screenshot of *MLB 2K7* and the Miller story below. There appear to be a very limited number of stories told in the game, however, and a single story is used repeatedly. Repetition takes the entertainment value out of the story as video game players do not wish to hear the same story multiple times. Once more stories are added to the story library, the problem of mapping game states to relevant stories becomes apparent. This is the problem we are studying in this thesis.

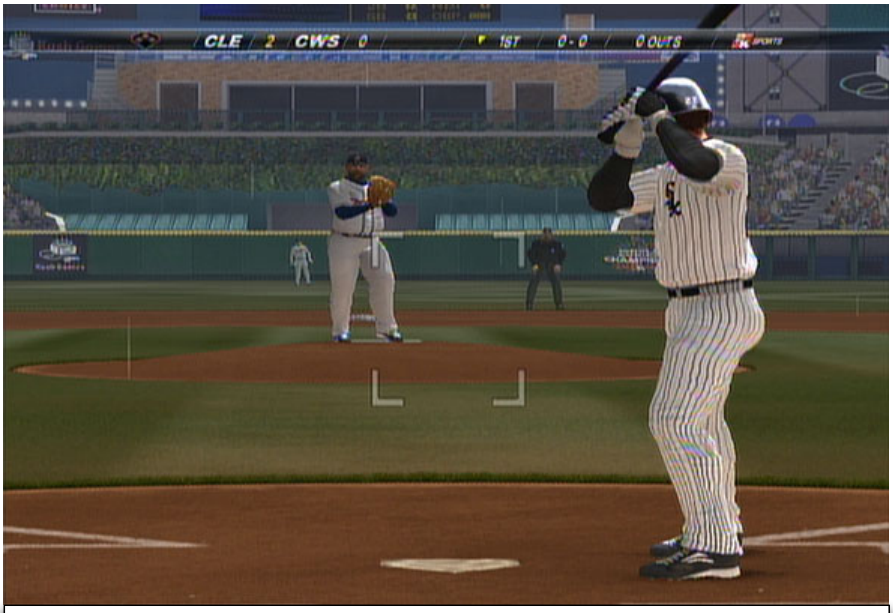

Figure 3.1: An example of storytelling in *MLB2K7*.

Robot World-Cup Soccer (RoboCup) is a research testbed involving robots playing soccer [Kitano *et al.*, 1997]. There is also a RoboCup simulation league, where the games are not physically played, but are simulated on a computer. Both the physical and simulation leagues provide researchers with a standard common testbed in which to evaluate their AI strategies for various goals. Previous academic work in automated commentary has focused

primarily on automated play-by-play commentary. *Byrne*, *Rocco* and *MIKE* [Andre *et al.*, 2000a] are three systems that produce automated play-by-play commentary for RoboCup simulator league games. The three systems obtain their data from the Soccer Server [Kitano *et al.*, 1997], which summarizes the gameplay's main features – the player locations and orientations, the ball location and the score of the game. Each system generates natural language templates, filling in player and team names where appropriate, then uses text-to-speech software to verbalize the derived commentary. Dynamic Engaging Intelligent Reporter Agent (*DEIRA*) is a similar system to *Byrne*, *Rocco*, and *MIKE*, as it performs the same task, but in the sport of horse racing.

Rocco, *MIKE*, *Byrne* and *DEIRA* are designed to provide accurate play-by-play commentary for RoboCup soccer and horse racing, and accomplish exactly that. There are some attempts within these systems to provide colour commentary, but none go as far as to try to incorporate storytelling. That is, these systems tackle a problem different from what our work tackles – they automate factual commentary with some bias added, but do not implement colour commentary via stories. Our system could be used in conjunction with each of these automated play-by-play systems to create fully automated commentary, featuring both play-by-play and colour.

Rocco (RoboCupCommentator) is template-based, so that when an event is to be communicated to the audience, the appropriate template is filled with the proper names of teams and players and then output. An example of this would be a template " BLANK shoots and scores! BLANK now has the lead!" being filled with the player name "Smith" and the team name "New York" to become "Smith shoots and scores! New York now has the lead!". Available time, bias and report style are factors considered when choosing a template to output. Repetition is avoided by storing a history of previously uttered sentences. This also avoids repeating a player's name too often, where a pronoun could be used instead (i.e., "he" instead of "Smith"). *Rocco* keeps a queue of possible outputs, along with a saliency value for each of these. As a possible output grows in age, its saliency shrinks, because it becomes less and less relevant. If the system deems it has nothing of interest in its queue (i.e, all the saliency values are below the output threshold), background information (such as the names of the teams involved in the game) is delivered to the user instead.

Rocco 2 [Andre *et al.*, 2000b] replicates *Rocco*, but also involves two simulated spectators as commentators, who are fans of one of the teams involved. This adds some biases to the commentators, as they use language that expresses emotion towards the teams involved. When the team for which the spectators are cheering performs well, they describe the ac-

tion in a positive manner, whereas if the opposition does well, they report the action in a negative manner. Thus, there is some colour in *Rocco 2*, but it is along the lines of creating a "hostile" environment, as in [Bryant *et al.*, 1982]. The template-based architecture of both *Rocco* and *Rocco 2* is designed for play-by-play commentary, so colour commentary within these systems would likely need to come from an outside source, such as our system.

Byrne is similar to *Rocco*, but also includes a human face that changes its expression based on built-in biases and what is happening on the simulated field. Where *Byrne* differs from *Rocco* is in its addition of emotion to the comments (similar to *Rocco 2*). The human face has static characteristics such as a nationality and liking or disliking particular teams. These characteristics are part of the *emotion-generation module* used in *Byrne* to determine what emotions to attach to a given output. Each emotion contains a type (i.e., "sad"), a score (an intensity value for the emotion), a decay function (the emotion diminishes with time), a cause (what happened to cause the emotion) and a target (which is optional – where the emotion is directed). Similarly, (*DEIRA*) [Knoppel *et al.*, 2008] involves a virtual agent providing play-by-play commentary with emotion for virtual horse races. The agent in *DEIRA*'s expression and tone change according to what is happening on the field of play (or racetrack). *Byrne* and *DEIRA* are designed to add emotion to play-by-play commentary, but not to produce colour commentary. While not done now, the human face in *Byrne* could be used in conjunction with SCoReS to add colour commentary to *Byrne*, giving it stories told with emotion.

MIKE (Multiagent Interactions Knowledgeably Explained) makes use of six *Soccer Analyzer Modules* that perform different operations on the available data, communicate with each other, and make suggestions to the proposition pool, which leads to natural language output. There are three lower-level modules – *basic*, *shoot* and *technique* and three higher-level modules – *bigram*, *Voronoi*, and *statistic*.

The bigram module models ball movement as a Markov chain, using a 24×24 transition matrix that follows ball play, as there are 22 players (11 versus 11) and two goals. The matrix allows for analysis of pass success rates, number of shots and other statistics related to ball movement. As its name would lead one to believe, the Voronoi module calculates Voronoi diagrams for each team. This determines the defensive coverage and other positioning information, so that it can be remarked upon by *MIKE*.

There are six types of remarks used in *MIKE*: Explanation of Complex Events (higher-level changes in action), Evaluation of Team Plays (criticism of positioning and ball movement as it pertains to a team as a whole), Suggestions for Improving Play (advice for how to

help improve play), Predictions (what the system thinks will happen), Set Pieces (plays from dead ball situations such as throw-ins) and Passwork (criticism of ball movement). Commentary from *MIKE* can involve one or more of these types. Some of these (most notably "Evaluation of Team Plays") are more colour commentary than play-by-play commentary, but are not in any way storytelling. The statistics gathered by *MIKE*'s six modules could be used to suggest stories, however, giving our system a richer set of game state features.

Within its live online game summaries, Major League Baseball uses a system called *SCOUT* [Kory, 2012] that provides textual analysis of the current game. The viewer is shown information such as the type of pitches thrown during an at-bat and the tendencies of the batter with respect to the pitcher. While *SCOUT* provides some colour, it is mostly a statistical summary and currently does not tell stories. Our system could extend *SCOUT* by adding stories to MLB online game summaries.

Rhodes [Rhodes *et al.*, 2010] describes a system that follows Freytag's pyramid [Freytag and MacEwan, 1908] and adds dramatic commentary to sports. His system stores values for the different Bryant motifs [Bryant *et al.*, 1977], and a vocabulary for each that changes as their level of intensity increases. Each theme has a set of actions that can occur in the game that either increase its intensity (Freytag's rising action) or decrease its intensity (Freytag's falling action). One theme within the system is "Urgency", with some level 1 phrases being "looking shaky" and "fortune not on their side", and level 3 phrases being "doomed" and "beyond salvation". Lexicalised Tree-Adjoining Grammar is used to generate comments on the fly. While drama is added to the commentary with this program, it is an augmentation to play-by-play commentary moreso than an addition of colour commentary. Our system could further augment the dramatic commentary by adding story selection.

3.3 Information Retrieval Methods

As we will describe in Chapter 4, we frame the problem of automated story selection in sports commentary as an information retrieval (IR) problem. Information retrieval involves finding material (documents) that satisfies an information need [Manning *et al.*, 2008]. In the context of this work, an IR problem involves a user submitting a *query* in order obtain a set of *documents* relevant to the query. Each query-document pair is typically assigned a *match quality* on a discrete scale from 0 to 4 as shown in Table 3.1. Information retrieval algorithms attempt to sort documents according to their match quality for a query. This is called the *ranking problem*. IR algorithms are typically divided into three groups – point-

Match Quality	Appropriateness
4	Perfect
3	Very Good
2	Good
1	Poor
0	Completely Inappropriate

Table 3.1: The meaning of match quality values, in terms of how appropriate the given story is for the given game state.

wise, pairwise and listwise. In this section, we discuss each of these strategies in detail, citing relevant examples.

3.3.1 Pointwise Algorithms

Pointwise algorithms approximate the ranking problem as a classification or regression problem. These algorithms thus attempt to minimize classification and regression error, with each query-document pair's match quality being equally important in evaluation. In their most basic form, a classification or regression algorithm is used to estimate each query-document match quality and documents are ranked according to these estimate match qualities for each query. Decision trees, decision stumps, artificial neural networks and naïve Bayes classifiers are some examples of classification and regression techniques that can solve the ranking problem in this fashion [Mitchell, 1997].

McRank [Li *et al.*, 2007] is a pointwise algorithm that casts the ranking problem as a multiple classification (hence 'Mc') and multiple ordinal classification problem. Multiple ordinal classification takes advantage of the fact that there is an ordering in IR class labels (i.e., a match quality of 3 is better than a match quality of 2, see Table 3.1). Thus, McRank calculates the probability that a given query-document pair is greater than a given threshold in order to sort documents for a query. A gradient boosting tree algorithm is used to learn class probabilities for each query-document pair, and these class probabilities are converted to ranking scores using their expected relevance. Normalized Discounted Cumulative Gain (NDCG), a common IR evaluation metric described in Section 4.5, is upper-bounded with the muitlple-class classification error.

As pointwise algorithms use classification and regression error in learning a ranking function, they are solving the "wrong" problem (and often a harder problem), since they are later evaluated with a different metric – an IR metric. IR metrics tend to focus on the top of a ranked list (since users typically care more about the top-ranked documents than

documents further down the list) and pointwise algorithms treat all documents as equally important, no matter where they place in a ranked list.

3.3.2 Pairwise Algorithms

Like pointwise algorithms, pairwise algorithms reduce the ranking problem to a classification problem. Instead of minimizing the error on each query-document pair, however, pairwise algorithms learn a binary classifier to estimate which of a pair of documents is better for a query. The goal of pairwise algorithms is to minimize the number of inversions necessary in the current ranking of documents for a query.

RankNet [Burges *et al.*, 2005] is a pairwise algorithm that uses an artificial neural network to minimize a cross-entropy loss function. This loss function is a function of the difference between the system's output for each member of a pair of documents for a given query. The authors claim that casting ranking as an ordinal regression problem solves an unnecessarily hard problem, highlighting the fact that users typically only care about the relative position of a document in the ranked list for a query, not the (arbitrarily valued) estimated match quality of the query-document pair. RankNet's evaluation is carried out on single documents, outputting a relevance score for each document with respect to unseen queries. Fidelity Rank (FRank) [Tsai *et al.*, 2007] is similar to RankNet, but uses fidelity as a loss function instead of cross-entropy loss. This is because while cross-entropy is convex it is not bounded, whereas fidelity is bounded between 0 and 1 (but not convex).

RankBoost [Yoav Freund and Singer, 2003] is a boosting pairwise algorithm based on AdaBoost [Freund and Schapire, 1995] that combines various user rankings into one ranking, with the goal to produce a linear ordering of the set of documents by combining ranking features. Ranking features are a set of given linear orderings. SSRankBoost [Amini *et al.*, 2008] is an implementation of RankBoost using partially labeled data for bipartite ranking. MPBoost [Esuli *et al.*, 2006] preserves the magnitude of match qualities with the thinking that an algorithm should focus on documents that are far apart in match quality for a given query.

Both pairwise algorithms and pointwise algorithms make use of loss functions that are only somewhat related to the IR metrics they will be tested on. They focus their learning on query-document pair match qualities, which leads to a lack of focus on the top of a ranked list. For most applications, including ours, the match quality of the top-ranked documents is much more important than the match quality of documents further down the ranked list.

3.3.3 Listwise Algorithms

Listwise algorithms rank documents for queries by directly optimizing IR scoring metrics, and thus operating directly on a ranked list of documents instead of query-document pairs. Optimizing IR scoring metrics is difficult because they are neither smooth nor differentiable (making gradient descent difficult, for example). As such, additional steps must be taken to incorporate them into IR algorithms.

SVM-MAP uses support vector machines to optimize a relaxed version of Mean Average Precision (MAP), an IR metric described in Section 4.3.2. Support vector machines have been used in pairwise algorithms (Ranking SVM [Herbrich *et al.*, 2000]), and SVM-MAP has been shown to outperform these algorithms. SoftRank [Taylor *et al.*, 2008] approximates the evaluation measure, allowing it to be both smooth and differentiable. It avoids the sorting problem by treating scores as random variables, operating in a four step process. First, a score distribution is calculated, and then it is mapped to a rank distribution. SoftRank then approximates NDCG (Section 4.5) with an expected smoothed NDCG called SoftNDCG. Since SoftNDCG is smooth and differentiable, SoftRank then uses gradient descent to optimize it. In AdaRank [Xu and Li, 2007], IR metrics are embedded directly into an existing method (AdaBoost [Freund and Schapire, 1995]). AdaRank is described in detail in Section 4.4.1.

LambdaRank [Burges *et al.*, 2006] builds on RankNet, using artificial neural networks for ranking, but instead of minimizing cross-entropy loss, LambdaRank computes its gradients after the documents have been sorted by their match qualities with respect to a query, which allows it to optimize IR metrics directly. LambdaMART [Burges, 2010] is a boosted version of LambdaRank, combining LambdaRank with MART. MART [Friedman, 1999] is a boosted tree model, outputting a linear combination of regression trees. In LambdaMART, MART is used to model derivatives, while LambdaRank specifies derivatives during training. LambdaMART and LambdaRank can also be combined to form an IR algorithm, as we will discuss in the next section.

While listwise algorithms directly (or approximately) optimize IR metrics by operating on ranked lists instead of query-document pairs and are often shown to outperform pointwise and pairwise algorithms, they often do not output a match quality for the documents in the ranked list. In some applications (such as ours), this match quality is an important bit of information, necessary to decide the output of the system. Thus, it is sometimes necessary to combine listwise algorithms with pointwise or pairwise algorthms, as we discuss next.

3.3.4 Hybrid Methods

In order to utilize the advantages of listwise algorithms (operation on ranked lists) and pointwise algorithms (estimating match qualities for query-document pairs), several hybrid IR algorithms have been developed. IntervalRank [Moon *et al.*, 2010] uses isotonic regression with an implicitly defined loss function. This loss function is primarily listwise, but uses pairwise constraints and an optional penalty for incorrect pointwise scores (incorrect query-document match qualities). Boltzrank [Volkovs and Zemel, 1999] also makes use of pairwise information within a listwise approach. A conditional probability distribution over a ranked list of documents for a query is created in order to permit gradient ascent to be used in the evaluation metric. These probabilities take the form of a Boltzmann distribution (hence, "Boltzrank") based on an energy function that depends on both pointwise and pairwise potentials.

A linear combination of ranking models has been used to construct a ranking algorithm [Burges *et al.*, 2011], combining twelve ranking models in a linear fashion – eight bootstrap aggregating LambdaMART boosted tree models, two LambdaRank artificial neural networks and two MART models using logical regression cost (pointwise). This IR algorithm combination won the 2011 Yahoo! Learning to Rank Challenge, demonstrating the effectiveness of hybrid IR methods.

Chapter 4

Proposed Approach

In this chapter, we present an AI approach to solving the problem of delivering story-based colour commentary to a live baseball game. We start by framing the problem as an information retrieval problem (Section 4.1). We then describe the machine-learning techniques we used (Sections 4.2 – 4.4) and finally combine these techniques in Section 4.5.

4.1 Information Retrieval Framework

We approach the problem of automated story selection in sports as an information retrieval problem. In the context of sports, the game state for which we are seeking an appropriate story is treated as the query, while the candidate stories returned by the system are the documents. Our system's goal then, is given a game state, to return to the user a ranked list of stories, based on how appropriate they are for the game state. We assume that the game state is available live, such as is the case for Major League Baseball [Live XML game summaries, 2011]. We also assume that a database of stories has previously been collected.

Thus, the problem is to retrieve stories most appropriate for game states during live sports broadcasts. Once a story database has been obtained, a system must learn to match the stories to game states. The broader the story database, the more likely an appropriate story can be found that matches any given game state. As "being interesting" is an informal and subjective measure, we evaluate the quality of the mapping by incorporating the selected stories into a simulated broadcast and test the enjoyment of viewers and the interest of professional commentators.

The match quality $D(\vec{g}, \vec{s})$ between a game state $\vec{g} = (g_1, g_2, \ldots g_n)$ and story $\vec{s} = (s_1, s_2, \ldots s_p)$ is an integer on the 5-point scale from 0 for a completely inappropriate match to 4 for a perfect match (as seen in Table 3.1). Thus, the problem is given a game state $\vec{g}$, to retrieve a story $\vec{s}$ of the highest match quality $D(\vec{g}, \vec{s})$.

4.2 Training Data

We solve this problem by using IR techniques and machine learning. Rather than feed the game state and story features to IR algorithms, we made the connection between corresponding features more explicit. A *similarity vector* $\vec{c}$ was computed for a game state specified by feature vector $\vec{g}$ and a story specified by feature vector $\vec{s}$. Each component of vector $\vec{c}$ is the result of comparing one or more features of $\vec{g}$ to one or more relevant features of $\vec{s}$. To compare binary features we use the logical connectives. For instance, if we want to match the runner on first base features, then we take a biconditional over the corresponding features: $c = g \leftrightarrow s = 1 \leftrightarrow 1 = 1$. Figure 4.1 shows corresponding features of $\vec{g}$ and $\vec{s}$ being compared to set the value of c. For non-binary features we use feature-specific functions. For instance, if we compare the current inning number g and the inning number involved in a story s, then the similarity feature c is calculated as $(8 - |g - s|)/8$, where values closer to 1 indicate a closer pairing of inning features. Another example is the marquee matchup feature valued between 0 and 1. It indicates how well the story and game state match in terms of a strong hitter and a strong pitcher being involved. It is calculated by combining several statistical features for the batter and pitcher in $\vec{g}$ with a story category feature in $\vec{s}$. The full calculation of the marquee matchup feature is given in Appendix C.

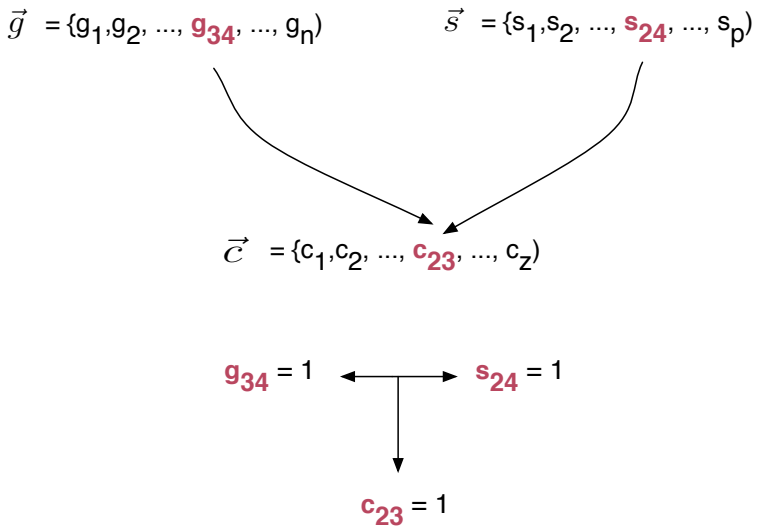

Figure 4.1: The runner on first base feature for the game state, g_{34} is compared to the corresponding feature in the story, s_{24} to produce a value for the runner on first base match feature in $\vec{c}$, c_{23}. If both g_{34} and s_{24} are 1, $c_{23} = 1$. Otherwise $c_{23} = 0$.

The similarity vector $\vec{c}$ indicates how related $\vec{g}$ and $\vec{s}$ are, but does not provide a scalar

value. We now need to map $\vec{c}$ to $D(\vec{g}, \vec{s})$ — the 5-point-scale quality of the match between $\vec{g}$ and $\vec{s}$. We use machine learning techniques to create this mapping from training data $\mathbf{T}$.

To build this training data set, we take a set of m game states vectors $\mathbf{G} = \{\vec{g}_1, \ldots, \vec{g}_m\}$ and form a set of $m \cdot p$ similarity vectors $\vec{c}$ for all $\vec{g}_i \in \mathbf{G}$ and all p stories vectors from our story vector library $\mathbf{S} = \{\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_p\}$. We then label each similarity vector with the ground-truth value of the quality of match between the corresponding game state and the story. Mathematically: $\mathbf{T} = \{(\vec{c}, D(\vec{g}, \vec{s})) \mid \vec{g} \in \mathbf{G}, \vec{s} \in \mathbf{S}, \vec{c}$ is the similarity vector for $\vec{s}$ and $\vec{g}\}$. For simplicity's sake in the rest of the thesis, we refer to $\mathbf{S}$ as the story library rather than the story vector library and $\mathbf{G}$ as the game state library, rather than the game state vector library. Also, as game states and stories are the equivalent of queries and documents in this work, we will use the former terms in the remainder of the thesis.

4.3 Listwise Scoring Metrics

IR algorithms are generally divided into three groups – pointwise, pairwise and listwise [Liu *et al.*, 2008]. Pointwise algorithms are regression and classification algorithms, with mean-squared error typically used as an error function. Pairwise algorithms perform a partial ordering on the data. Listwise algorithms make direct use of IR scoring metrics to search for a good ranking of stories. All IR algorithms output a ranker, which produces a ranked list of stories based on a given game state. The optimal ranker produces the list of stories ranked according to their true match qualities, $D(\vec{g}, \vec{s})$.

We used listwise algorithms to map game states to stories, as they rank stories for each game state and the ranking is evaluated with IR scoring metrics. These metrics generally focus on the top of a ranked list, giving more importance to the match quality of the top ranked stories and decreasing the importance of accuracy with the ranking of stories.

The process of evaluating a ranker using IR metrics is shown in Figure 4.2. A ranker R sorts a story database $\mathbf{S}$ based on the current game state $\vec{g}$. This permutation (sorting) of $\mathbf{S}$, π, is then scored by D to produce a vector of match qualities θ. The metric M then maps θ to a scalar value, providing feedback about the strength of the ranking.

Formally, a ranker performs a mapping of stories: $R_{\mathbf{S}} : \mathbf{G} \rightarrow \Pi$ from the set of game states $\mathbf{G}$, to a set of permutations Π, for a given set $\mathbf{S}$ of p stories. A given permutation of $\mathbf{S}$, $\pi \in \Pi$ is scored according to its match qualities with $\vec{g}$, D, producing θ. Thus, θ is vector of p integers, $\theta_i = D(\pi_i, \vec{g})$. IR scoring metrics accept this θ as input and output feedback γ on the quality of the permutation (ranking) π. Formally, an IR metric performs a mapping

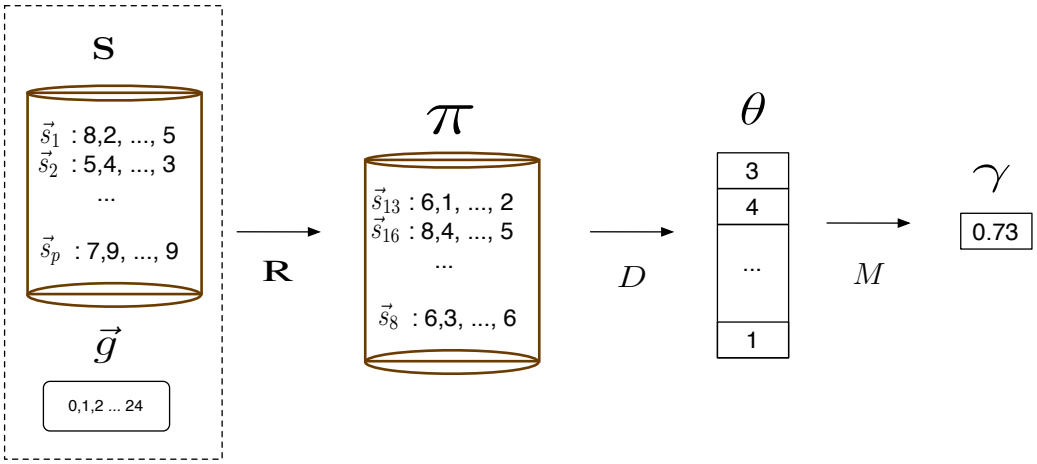

Figure 4.2: Evaluating a ranker.

Winner Takes All (θ, t)

INPUT:

θ	vector of match qualities for a ranking π
t	threshold

OUTPUT:

γ:	score for θ

```
1    if θ_1 ≥ t
2        γ ← 1
3    else
4        γ ← 0
5    end if
```

Figure 4.3: Winner Takes All scoring metric.

$M : \mathbb{Z}^p \rightarrow \mathbb{R}$. Some IR scoring metrics also require the match qualities θ^*, produced by the ground truth ordering π^*.

4.3.1 Winner Takes All

A simple example of an IR scoring metric is Winner Takes All (WTA), shown in Figure 4.3. This metric considers only the match quality for the top ranked story, θ_1 comparing it to a given threshold t. If the top ranked story's match quality meets or surpasses the threshold (line 1), WTA returns 1, otherwise, it returns 0 (lines 2-5). Thus, WTA is a binary threshold scoring metric whose output rests completely on the top ranked story, ignoring the rest of the ranking. Consider the example in Table 4.1. Here $\theta_1 = 2$. If the threshold $t = 2$, then WTA would return 1. If $t > 2$, however, WTA would return 0.

Rank	θ	θ^*
1	2	4
2	4	4
3	1	3
4	3	3
5	2	2

Table 4.1: An example of the scores θ for a ranking π, for a given game state $\vec{g}$ and the best possible scores, θ^* for optimal ranker π^*, for the top 5 ranking positions.

4.3.2 Average Precision

Other IR scoring metrics consider more stories than the top ranked story. One example is Average Precision (AP), shown in Figure 4.4. AP accepts the same input as WTA, but also requires the number of positions to consider relevant, N. That is, AP requires the user to specify how much of the ranked list is important. In our context, N varies depending upon the application. If we would like to suggest stories to professional commentators, then N can be 3 or 4, to give the commentators some choice. When operating autonomously, however, N is 1, as only the top story will be output to the viewer.

Average Precision (θ, t, N)

INPUT:

θ	vector of match qualities for a ranking π
t	threshold
N	number of ranking positions to consider

OUTPUT:

γ:	score for θ

```
1    γ ← 0
2    for i = 1, ..., N
3        if θ_i ≥ t
4            R(i) ← 1
5        else
6            R(i) ← 0
7        end if
8        P(i) ← (Σ_{j=1}^{i} R(j))/i
9    end for
10   for i = 1, ..., N
11       if R(i) = 1
12           γ ← γ + P(i)
13       end if
14   end for
15   γ ← γ / Σ_{i=1}^{N} R(i)
```

Figure 4.4: Average Precision scoring metric.

At each position of θ to be considered, AP checks if θ_i meets the threshold t (line 3). If so, this position is considered relevant (line 4). The precision P at any position i is the total number of relevant stories in the top i positions, divided by i (line 8). The average precision, γ, is then calculated, based on the precision at each relevant position (line 12) and the total number of relevant positions (line 15). Higher values of γ suggest a better ranking, in the range $[0, 1]$. Consider again the example in Table 4.1, and let $N = 4$ and $t = 3$. The AP here would be calculated as:

$$\gamma = \frac{1}{2}\left(\frac{1}{2} + \frac{2}{4}\right) = 0.5.$$

If we instead let $t = 2$, then,

$$\gamma = \frac{1}{3}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{4}\right) \approx 0.92.$$

Normalized Discounted Cumulative Gain(θ, N, θ^*)

INPUT:

θ	vector of match qualities for a ranking π
N	number of ranking positions to consider
θ^*	ground truth ranking match qualities

OUTPUT:

γ:	score for θ

```
1   E ← 0
2   V ← 0
3   for i = 1, ..., N
4       E ← E + (2^θ_i − 1)/log_2(i + 1)
5       V ← V + (2^θ*_i − 1)/log_2(i + 1)
6   end for
7   γ = E/V
```

Figure 4.5: Normalized Discounted Cumulative Gain scoring metric.

4.3.3 Normalized Discounted Cumulative Gain

Another example of an IR scoring metric is Normalized Discounted Cumulative Gain (NDCG), shown in Figure 4.5. We will demonstrate NDCG from the bottom up. The gain at position i is calculated using the match quality at this position ($2^{\theta_i} - 1$). The cumulative gain is the sum of these gains over all positions N to be considered. The discount factor is $1/\log(i + 1)$, increasing the discount as the rank position worsens. Finally, the metric is normalized by dividing by the optimal Discounted Cumulative Gain – that of the ground truth scores θ^* for the given $\vec{g}$.

NDCG takes as input the vector of match qualities θ for a ranking π, the number N of positions to consider, and the ground truth match qualities π^* for the given game $\vec{g}$. First, the true DCG, E, and the optimal DCG, V, are initialized to 0 (line 1). Then for every position i up to N, E and V are updated by adding the Discounted Gain at i (lines 3 – 6). Finally, the NDCG is calculated by dividing the true DCG by the optimal DCG (line 7). As with AP, NDCG scores have the range $[0, 1]$, with higher scores suggesting a better ranking.

Consider the case where $N = 1$ for the example in Table 4.1:

$$\gamma = \frac{(2^2 - 1)}{\log_2(1 + 1)} \Big/ \frac{(2^4 - 1)}{\log_2(1 + 1)} = \frac{3}{15} = 0.2.$$

If $N = 2$:

$$\gamma = \Big(3 + \frac{(2^4 - 1)}{\log_2(2 + 1)}\Big) \Big/ \Big(15 + \frac{(2^4 - 1)}{\log_2(2 + 1)}\Big) \approx \frac{12.46}{24.46} \approx 0.51.$$

And if $N = 3$:

$$\gamma = \Big(12.46 + \frac{(2^1 - 1)}{\log_2(3 + 1)}\Big) \Big/ \Big(24.46 + \frac{(2^3 - 1)}{\log_2(3 + 1)}\Big) \approx \frac{12.96}{27.96} \approx 0.46.$$

4.3.4 Discussion

Each of the described IR metrics focuses on the top of a ranked list. That is, the match qualities at the top of θ have a greater bearing on the score provided by an IR metric than those at the bottom of θ. WTA is the extreme case, considering only whether the match quality of the top ranked story meets a given threshold. This makes WTA seem well-suited to the case of an autonomous commentator, as stories ranked second or worse are irrelevant for its output. WTA provides little information during training, however, making it possibly difficult for an IR algorithm to effectively learn with WTA as its scoring metric.

AP can consider more than one position (depending on N), but is again binary (either a story is relevant, or it is not). NDCG makes use of both the actual match qualities of θ in its gain calculation, as well as the best possible ranking for the given $\vec{g}$. AP ignores both these bits of information.

Table 4.2 shows how WTA, AP and NDCG vary with N and t, given the data in Table 4.1. WTA does not vary with N and will output 1 if $t \leq 2$ and 0 otherwise. Note that AP is unaffected by any story s_i it considers irrelevant unless there is a story s_j ranked lower than s_i that is relevant. NDCG, on the other hand, will have its score decrease with less relevant stories, even at θ_N (although at a discounted rate as we approach N). Thus NDCG is

N	**NDCG**	**AP** ($t = 2$)	**AP** ($t = 3$)	**WTA**($t = 2$)	**WTA** ($t = 3$)
1	0.2	1	0	1	0
2	0.51	1	0.5	1	0
3	0.46	1	0.5	1	0
4	0.52	0.92	0.5	1	0
5	0.53	0.89	0.5	1	0

Table 4.2: How IR scoring metrics NDCG, AP and WTA vary with respect to N and t, given the data in Table 4.1.

more likely to "protect" a colour commentator from bad stories, whereas AP does not allow these poor stories to affect the output of better stories at the top of θ. NDCG gives a finer evaluation to θ as it uses more information than AP (namely, θ^*), possibly providing an IR algorithm more feedback with respect to its rankings. As each of WTA, AP and NDCG offer different advantages and disadvantages, we considered each in our empirical work.

Expected Reciprocal Rank (ERR) [Chappelle *et al.*, 2009] is another common IR scoring metric similar to NDCG, but whose discount factor at a position varies according to the relevance of the story in the position before (similar to AP). Return Score (RS) is a metric we created that is the same as WTA, except that it returns the actual match quality of the top ranked story, rather than simply perform a threshold operation on it. We considered both ERR and RS in our preliminary experiments, but neither was chosen to be part of a final ranker, so we omit further discussion of them here.

4.4 SCoReS Offline

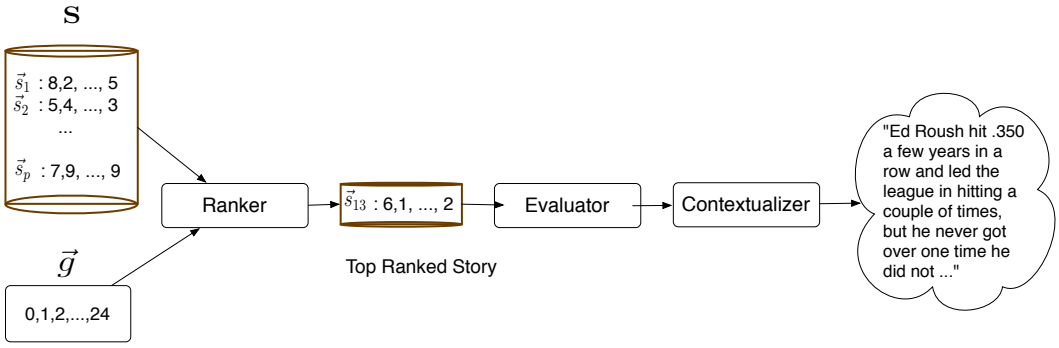

Figure 4.6: SCoReS chooses a story to output to a commentator.

Our approach to selecting stories for game states is a three-step process involving a *Ranker*, an *Evaluator*, and a *Contextualizer*, as shown in Figure 4.6. This hybrid approach

uses a listwise algorithm to rank, a pointwise algorithm to evaluate and domain-specific methods to contextualize. In this section, we describe the offline training and creation of each element of SCoReS, and then describe the online operation in detail (Section 4.5).

4.4.1 Machine-learning a Ranker

For the ranking algorithm, we adapted AdaRank [Xu and Li, 2007], a listwise algorithm based on AdaBoost [Freund and Schapire, 1995]. AdaRank forms a "strong" ranker by iteratively selecting and combining "weak" rankers. A weak ranker uses a single component of the similarity vector $\vec{c}$ to rank (i.e., sort) the training data $\mathbf{T}$. Each weak ranker has a "vote" on the final ranking, based on how its mapping of $\mathbf{S}$ to π over the training data was scored according to a chosen IR scoring metric (described below).

SCoReS AdaRank (Algorithm **??**) accepts as input a set of training data $\mathbf{T}$, the number of game states in $\mathbf{T}$, m, an IR scoring function M, the number of weak rankers to compose the strong ranker k, and the number of tie-breaking features to use y. In line 1, SCoReS AdaRank first partitions $\mathbf{T}$ by its m constituent games states, where i runs from 1 to m. This is done because stories can be meaningfully sorted by the match quality values (D) only for a given game state. The ground truth rankings $\mathcal{T}$ are then calculated (line 2) for possible use in evaluating weak rankers (line 13). All weights are initialized to $1/m$ (line 3) as all game states are initially equally important. The main ranker R and its corresponding confidence values A are initialized to be empty sets (line 4). The set of feature combinations to be considered for use in weak rankers, B, is then calculated based on the number of features in each $\vec{c}$ in $\mathbf{T}$, and the number of features to use for tie-breaking y (line 5).

At each iteration of SCoReS AdaRank, elements b of B whose first elements have not yet been used in R are considered as possible weak rankers (lines 6-25). Thus, each feature of $\vec{c}$ may only be used once as the main sorter in a weak ranker. For each game state, the weighted score v of sorting $\mathbf{T}_i$ by b (with any remaining ties broken randomly) is calculated, using the scoring function M and current weights w (lines 10-14). The arguments to M vary based on which scoring metric is used, but all metrics we consider accept θ_i, the match qualities for T_i as input. If the mean weighted score v for b is greater than the maximum encountered so far in this iteration, then the feature combination to be used in the weak ranker r for this iteration is set to b (lines 15-18). After evaluating all valid elements of B, the best weak ranker for this iteration is added to the main ranker R (line 21) and A and w are updated (lines 22-24) as in [Xu and Li, 2007]. The data is re-weighted after each iteration of SCoReS AdaRank, so that examples that have been incorrectly classified so far

SCoReS AdaRank ($\mathbf{T}, m, M, k, y$)

INPUT:

$\mathbf{T}$:	training data
m	number of game states
M:	IR scoring function
k:	number of iterations
y:	number of tie-breaking features

OUTPUT:

R:	ranker

partition $\mathbf{T}$ by game state: $\mathbf{T} = \mathbf{T}_1 \cup \mathbf{T}_2 \cup \ldots \cup \mathbf{T}_m$
sort each $\mathbf{T}_i$ by D values, yielding ground truths $\mathcal{T}_i$
initialize weights $w(\mathbf{G})$ to $1/m$
initialize ranker R and weak ranker confidences A to $\emptyset$
get all combinations B of length $y + 1$ from $\mathbf{T}$
for each iteration up to k
    $\mu \leftarrow 0$
    for each combination b in B
        if $b(1) \notin R$
            for $i = 1, \ldots, m$
                sort $\mathbf{T}_i$ by b, yielding $\mathbf{T}'_i$
                $\theta_i \leftarrow D(\mathbf{T}_i)$
                $v(i) \leftarrow w(i) \cdot M(\theta_i, \ldots)$
            end for
            if mean$(v) > \mu$
                $\mu \leftarrow$mean(v)
                $r \leftarrow b$
            end if
        end if
    end for
    add r to R
    calculate α for r
    add α to A
    update w
end for

Figure 4.7: SCoReS AdaRank algorithm.

Game State 1					$w_1 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	0	0.875	0.6	4
2	0	1	0.5	0.4	2
3	1	1	0	0	1
4	0	0	0.375	0.9	3
Game State 2					$w_2 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	0	0.5	0.7	0
2	1	1	0.5	0.7	3
3	0	1	0	0.3	2
4	1	0	0.125	0.5	4
Game State 3					$w_3 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	1	1	0.3	4
2	1	0	0	0.3	2
3	1	0	0.75	0.5	1
4	0	0	0.875	0	0

Table 4.3: Unordered training data $\mathbf{T}$ for SCoReS AdaRank.

are given more weight (line 24).

As a small, concrete example, let M be NDCG [Xu and Li, 2007], $m = 3$, $k = 2$, and $y = 1$. Let the similarity vectors $\vec{c}$ in $\mathbf{T}$ consist of the following features: the runner on first base feature (c_1), the strikeout feature (c_2), the inning feature (c_3), and the marquee matchup feature (c_4). The first two are binary: they are 1 if both the story and game state involve a runner on first base (or both involve a strikeout), and 0 otherwise. The marquee matchup and inning features are computed as previously described. Table 4.3 shows possible training data split into its 3 constituent game states, while Table 4.4 shows the ground truth orderings, π^*. The weight of each game is initially set to $1/3$. The best feature combination for the first iteration is (c_3, c_4) — the runner on first base feature to be used as the main sorter, with the strikeout feature as a tiebreaker. Sorting by this feature combination yields the orderings shown in Table 4.5. Assuming we consider the top 3 ranking positions relevant, the NDCG scores for each game state in this ordering would be $(0.97, 0.59, 0.89)$, with weighted mean $\mu = 0.81$ (as all weights are equal).

The weak ranker r for this iteration would thus be (c_3, c_4) and c_3 would not be considered as a main sorter in weak rankers in later iterations. α is calculated with the formula (from [Xu and Li, 2007]):

Game State 1					$w_1 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	0	0.875	0.6	4
4	0	0	0.375	0.9	3
2	0	1	0.5	0.4	2
3	1	1	0	0	1
Game State 2					$w_2 = 1/3$
Story	c_1	c_2	c_3	c_4	D
4	1	0	0.125	0.5	4
2	1	1	0.5	0.7	3
3	0	1	0	0.3	2
1	0	0	0.5	0.7	0
Game State 3					$w_3 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	1	1	0.3	4
2	1	0	0	0.3	2
3	1	0	0.75	0.5	1
4	0	0	0.875	0	0

Table 4.4: Ground truth ordering π^* of training data $\mathbf{T}$ from Table 4.3.

Game State 1					$w_1 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	0	0.875	0.6	4
2	0	1	0.5	0.4	2
4	0	0	0.375	0.9	3
3	1	1	0	0	1
Game State 2					$w_2 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	0	0.5	0.7	0
2	1	1	0.5	0.7	3
4	1	0	0.125	0.5	4
3	0	1	0	0.3	2
Game State 3					$w_3 = 1/3$
Story	c_1	c_2	c_3	c_4	D
1	0	1	1	0.3	4
4	0	0	0.875	0	0
3	1	0	0.75	0.5	1
2	1	0	0	0.3	2

Table 4.5: The ordering π of $\mathbf{T}$ after being sorted by feature combination (c_3, c_4). Random tie-breaking was necessary for game state 2 (between stories 1 and 2).

$$\alpha = \frac{1}{2} \cdot \ln \frac{\sum_{i=1}^{m} w(i)\,(1 + M(\theta_i, 3, \theta_i^*))}{\sum_{i=1}^{m} w(i)\,(1 - M(\theta_i, 3, \theta_i^*))}$$

where θ_i is the vector of match qualities D for the ordering of game state i by r, and θ_i^* is the vector of match qualities D for the ground truth π^*. α for this iteration would thus be 1.13. The weights for each game state are then updated with the formula (from [Xu and Li, 2007]):

$$w(i) = \frac{e^{-M(\theta_i, 3, \theta_i^*)}}{\sum_{j=1}^{m} e^{-M(\theta_i, 3, \theta_i^*)}}$$

giving the new $w = (0.28, 0.41, 0.32)$.

On the second iteration, game state 2 has more weight, as the weak (and main) ranker (c_3, c_4) failed to sort it as well as it did game states 1 and 3. This leads to $r = (c_1, c_4)$, $\mu = 0.73$, $\alpha = 1.01$, and thus main ranker $R = ((c_3, c_4), (c_1, c_4))$ and $A = (1.13, 1.01)$. After normalizing, $A = (0.53, 0.47)$.

This means the weak ranker using the inning feature to sort and the marquee matchup feature as a tiebreaker gets 53% of the vote for ranking $\mathbf{S}$, and the runner on first base feature combined with the marquee matchup feature tiebreaker gets 47% of the vote with respect to the ranking output by SCoReS AdaRank.

4.4.2 Machine-learning an Evaluator

Though the *Ranker* R output by SCoReS AdaRank provides a ranked list of stories, it does not provide a value for D for these stories. Thus, there is always a top-ranked story for a game state, but SCoReS AdaRank provides no indication as to how "good" the top ranked story is. We added an *Evaluator* to provide an estimate of D. The *Evaluator* can then be used as a threshold to ensure the top-ranked SCoReS AdaRank story is worth telling.

In principle, it is possible to use the *Evaluator* on its own to rank stories. The reason we do not do this is because of its pointwise nature; it uses mean squared error (MSE) as a scoring metric, which treats each datum equally – there is no preference with respect to accuracy towards the top of a ranked list. As a result, the *Evaluator* tends to give only a rough "good/no good" evaluation of a story and is not sensitive enough to choose between two similar stories. Thus, we instead use a pointwise algorithm to obtain a quick estimate of one particular match quality, but do not rely on it to estimate every match quality and rank the stories based on these estimates.

4.4.3 Coding a Contextualizer

Even if the *Ranker/Evaluator* combination deem a story appropriate for output given a game state, it may not be a good time to tell the story. This is because certain situations are not appropriate for stories as they are exciting on their own, or because the story may get cut off if we attempt to tell it. We use several domain specific features to determine whether it is a good time to tell a story. An example from baseball where a story may be cut off would be if there are two strikes on the batter with two out. This would be a bad time to start telling a story as the inning could easily end on the next pitch (with a strikeout). If the game at hand is the seventh game of the World Series, tied in the ninth inning with the bases loaded and two out, there is probably no need to tell a story, and telling one would likely aggravate viewers, who would be completely focused on the play.

SCoReS ($\mathbf{G}', \mathbf{S}, R, A, E, t$)

INPUT:

$\mathbf{G}'$:	game states for a live game
$\mathbf{S}$:	our story library
R:	*Ranker* from SCoReS AdaRank
A:	weak ranker confidences for R
E	*Evaluator*
t	threshold for *Evaluator*

```
1    for each game state g⃗ in G′
2        if should tell a story in g⃗
3            create {c⃗}, comparing g⃗ to s⃗ ∈ S
4            rank {c⃗} with R and A
5            s⃗_∘ ← top ranked story
6            if E(s⃗_∘) ≥ t
7                output s⃗_∘ to broadcast team (or viewer)
8            end if
9        end for
10   end for
```

Figure 4.8: SCoReS as used during a live game.

4.5 SCoReS Online

Our SCoReS system thus consists of a *Ranker* (learned by SCoReS AdaRank), an *Evaluator* (learned by a pointwise algorithm) and a *Contextualizer* (manually designed) as shown in Figure 4.8. SCoReS processes each game state in a given (novel) game $\mathbf{G}'$ (line 1). If the current game state $\vec{g}$ is appropriate for a story according to the *Contextualizer* (line 2), we create similarity vectors of $\vec{g}$ and each story in $\mathbf{S}$ (line 3). The similarity vectors are then sorted with the ranker R and confidences A learned by SCoReS AdaRank offline (line 4).

The top ranked story is extracted in line 5 and then scored by the *Evaluator* E in line 6. If the story passes the provided threshold t, it is suggested to the broadcast team (line 7). Or, if SCoReS is operating autonomously, the story is output to the viewer.

Chapter 5

Empirical Evaluation

To evaluate the quality of SCoReS, we conducted a series of empirical studies where we asked potential users to evaluate the system. We examined several possible applications of SCoReS: providing generic commentary, adding stories to existing commentary, picking context appropriate stories, and assisting professional commentators to select stories.

We conducted several experiments to evaluate whether SCoReS improved broadcast quality in any of these applications. User Study I was conducted to ensure commentary was beneficial within our game library before we even tried to improve said commentary. User Studies II – VI tested whether there was a reason to tell stories in a broadcast within our game library and whether the stories selected needed to be in proper context. The final experiment was a demonstration of the SCoReS system to professional commentators. Their feedback is an estimate of whether SCoReS can be successfully deployed in a professional broadcast setting.

These experiments evaluated SCoReS in both its modes of operation: as an autonomous commentary tool within the user studies, and as an assistant to a human colour commentator in the interviews with professional commentators. In the user studies, we evaluated commentary augmented by SCoReS by inserting SCoReS' top-ranked story into video clips from actual baseball games. We chose to use actual games because watching clips from a video game would likely not be interesting to the participants, inducing boredom. Note that improving commentary from actual games appears to be a harder problem than improving sports video game commentary, as commentary in video games is pre-recorded and generic, whereas commentary during an actual game is tailored to that particular game.

In order to build training data for SCoReS, we first downloaded MLB game statistics from MLB's XML site [Live XML game summaries, 2011], with permission from MLB. For all experiments, the game state and story features were kept constant. The game state

features used in our experiments are given in Table 5.1. The feature vector for each game state included the values for the above-described features at a particular pitch, as well as their values for the previous pitch. This was done because stories are often told in reference to an event preceding the current game state. Thus, g_3 stored the current number of outs for the current batter, while g_{40} stored the out total from the previous pitch.

One hundred and ten stories were gathered from "Rob Neyer's Big Book of Baseball Legends" [Neyer, 2008], "Baseball Eccentrics" [Lee and Prime, 2008], and Wikipedia. A selection of these stories is shown in Appendix B. Stories ranged in year from 1903 to 2005. Feature selection and categorization of the stories were done by hand. Using our baseball expertise, and based on the features available from MLB's XML site, we chose the 10 game state categories listed in Table 5.2. While these categories are not those used in other work, we thought they were appropriate given the available data (i.e., it would be difficult to categorize a game state as "the Lucky Break Victory" as in [Ryan, 1993] without video of the game,using only selected statistics). Story features used in all experiments are detailed in Table 5.3. Building similarity vectors $\vec{c}$ created a new feature set, given in Table 5.4.

In order to ease the burden on the human performing the match quality labelling, we constructed some rules based on the available statistics that suggest a match quality, which are then refined by hand. A screenshot of the suggestion program is shown in Figure 5.1. Relevant statistics from both the game (e.g., batting statistics) and the story (e.g., the year and category) are displayed. The title and first few words of the story are displayed for the labeller, to show which story is at hand, and allow for reference to the rest of the story statistics, stored elsewhere. Matching statistics are given at the bottom, which in this case are a team ("nya" is the New York Yankees), the month and a double play. The double play in the game refers to the result of the previous at bat, and in this case, the previous half inning. This can be gathered from the information presented, as there are 0 outs in this half inning, meaning the double play must have occurred in the previous half inning. The program suggests a score of 2 (meaning "Good") which can be refined by the labeller.

The suggested match qualities for each $\vec{c}$ are calculated based primarily on the confidence that the story and game state match on category $(1/3)$ and whether any of the teams matched $(1/3)$. The confidence of a categorical match between the game state and story is based upon features of the game state relevant to the story category (the calculation of the marquee matchup category match is given in Appendix C). For instance, if the story category is "bad statistics for batter", then the confidence in a category match is based on the statistics of the current batter. If the story category is "opening of inning", then the

id	Name	Type	Description
g_1	Balls	Integer	Number of balls on batter.
g_2	Strikes	Integer	Number of strikes on batter.
g_3	Outs	Integer	Number of outs in inning.
g_4	Batter Side	Binary	0 for right, 1 for left.
g_5	Pitcher Side	Binary	0 for right, 1 for left.
g_6	Season At Bats	Integer	Batter's number of at-bats this season.
g_7	Season Average	Percentage	Batter's batting average this season.
g_8	Season HR	Integer	Batter's home run total this season.
g_9	Season RBI	Integer	Batter's run batted in total this season.
g_{10}	Season Wins	Integer	Pitcher's win total for this season.
g_{11}	Season Strikeouts	Integer	Pitcher's strikeout total this season.
g_{12}	Season Pitcher Average	Percentage	Pitcher's batting average against this season.
g_{13}	Season Pitcher At Bats	Integer	Number of at bats versus pitcher this season.
g_{14}	Career At Bats	Integer	Batter's number of at-bats in his career.
g_{15}	Career Average	Percentage	Batter's batting average in his career.
g_{16}	Career HR	Integer	Batter's home run total in his career.
g_{17}	Career RBI	Integer	Batter's run batted in total in his career.
g_{18}	Career Wins	Integer	Pitcher's win total in his career.
g_{19}	Career Strikeouts	Integer	Pitcher's strikeout total in his career.
g_{20}	Career Pitcher Average	Percentage	Pitcher's batting average against in his career.
g_{21}	Career Pitcher At Bats	Integer	Number of at bats versus pitcher in his career.
g_{22}	Road Score	Integer	Number of runs scored by the road team.
g_{23}	Home Score	Integer	Number of runs scored by the home team.
g_{24}	Road Wins	Integer	Number of wins this season for the road team.
g_{25}	Road Losses	Integer	Number of losses this season for the road team.
g_{26}	Home Wins	Integer	Number of wins this season for the home team.
g_{27}	Home Losses	Integer	Number of losses this season for the home team.
g_{28}	Home Team	Integer	Home Team ID.
g_{29}	Road Team	Integer	Road Team ID.
g_{30}	Inning	Integer	Current Inning.
g_{31}	Last At-Bat	Integer	ID for result of last at bat (e.g., "pop out").
g_{32}	Month	Integer	Current Month.
g_{33}	Day	Integer	Current Day.
g_{34}	Runner on 1st	Binary	True if there is a runner on first base.
g_{35}	Runner on 2nd	Binary	True if there is a runner on second base.
g_{36}	Runner on 3rd	Binary	True if there is a runner on third base.
g_{37}	Pitch Count	Integer	Total number of pitches thrown by current pitcher.

Table 5.1: Game state features used in all experiments. There were actually 74 game state features, as each game state vector stores values for the previous game state as well.

id	Story Category
1	Opening of Inning
2	Bad Statistics for Batter
3	Marquee Matchup
4	Great Statistics for One Player
5	Bad Statistics for either Hitter or Pitcher
6	Important Games from History
7	Big Finish
8	Blowout or Comeback
9	One-run Game, Home Run Hitter Batting
10	Human Interest

Table 5.2: The story and game state categories used in our experiments.

confidence is based on how close the game state is to the beginning of an inning.

The result of the last at bat matching events in the story provides $1/18$ of the suggested match quality. The remaining portion of the match quality depends upon the sum of the following corresponding features (if they are true in both the game and the story): the day, the inning, the month, the run difference, the runner positions, the count on the batter, the number of balls, strikes and outs and whether the game state and the story involve a "blowout" (a difference of 5 or more runs). While many of these criteria resemble the similarity features in Table 5.4, they are not equivalent. The similarity features were designed after the data was labelled, and provide more information than the suggestion program.

5.1 Choosing a Ranker

In order to choose a *Ranker* and an *Evaluator* for SCoReS, we performed a leave-one-out cross validation experiment. Training data consisted of 40 game states and the 110 stories gathered from various sources. At each cross-validation fold, 4290 data (39 game states $\times$ 110 stories) were used to build training data $\mathbf{T}$, while 110 (1 game state $\times$ 110 stories) were used for testing data. The same story database was used in training and testing. Candidate *Evaluator*s were trained on $\mathbf{T}$ and then used within SCoReS at each fold. As a reminder, SCoReS AdaRank accepts as input a set of training data $\mathbf{T}$, the number of game states in $\mathbf{T}$, m, an IR scoring function M, the number of weak rankers to compose the strong ranker k, and the number of tie-breaking features to use y. While $\mathbf{T}$ and m are determined by the data, M, k and y must be otherwise determined.

id	Name	Type	Description
s_1	Month	Integer	Current Month.
s_2	Inning	Integer	Current Inning.
s_3	Run Difference	Integer	Difference in score.
s_4	Year	Integer	Year story took place.
s_5	Home Team	Integer	Home Team ID.
s_6	Road Team	Integer	Road Team ID.
s_7	Strikeout	Binary	True if story contained a strikeout.
s_8	Home Run	Binary	True if story contained a home run.
s_9	Sacrifice	Binary	True if story contained a sacrifice bunt or fly.
s_{10}	Double	Binary	True if story contained a double.
s_{11}	Triple	Binary	True if story contained a triple.
s_{12}	Double play	Binary	True if story contained a double play.
s_{13}	Fly Out	Binary	True if story contained a fly out.
s_{14}	Pop Out	Binary	True if story contained a pop out.
s_{15}	Ground Out	Binary	True if story contained a ground out.
s_{16}	Triple Play	Binary	True if story contained a triple play.
s_{17}	True or False	Binary	True if story is factually true.
s_{18}	Injury	Binary	True if story contained an injury.
s_{19}	Pinch Hitter	Binary	True if story contained a pinch hitter.
s_{20}	Rally	Binary	True if story contained a rally.
s_{21}	Blowout	Binary	True if story contained a blowout victory.
s_{22}	Substitution	Binary	True if story contained a substitution.
s_{23}	Foul Ball	Binary	True if story contained a foul ball.
s_{24}	Runner on 1st	Binary	True if story contained a runner on first base.
s_{25}	Runner on 2nd	Binary	True if story contained a runner on second base.
s_{26}	Runner on 3rd	Binary	True if story contained a runner on third base.
s_{27}	Outs	Integer	Number outs in the game featured in the story.
s_{28}	No Hitter	Binary	True if story contained a no hitter.
s_{29}	Walk	Binary	True if story contained a walk.
s_{30}	Intentional Walk	Binary	True if story contained an intentional walk.
s_{31}	Single	Binary	True if story contained a single.
s_{32}	Pitcher Home Run	Binary	True if story contained a pitcher home run.
s_{33}	Bunt	Binary	True if story contained a bunt.
s_{34}	Inside Park Home Run	Binary	True if story contained an inside the park home run.
s_{35}	Hit By Pitch	Binary	True if story contained a hit batsman.
s_{36}	Ejected	Binary	True if there was an ejection in the story.
s_{37}	World Series	Binary	True if current game is a World Series game.
s_{38}	Grand Slam	Binary	True if story contained a .grand slam.
s_{39}	Play at Plate	Binary	True if story contained a play at the plate.
s_{40}	Debut	Binary	True if story contained a player's MLB debut..
s_{41}	Balls	Integer	Number of balls on batter.
s_{42}	Strikes	Integer	Number of strikes on batter.
s_{43}	Assist	Binary	True if story contained an outfield assist.
s_{44}	Error	Binary	True if story contained a fielding error.
s_{45}	Category	Integer	ID for Category (Category list is given in 5.2).

Table 5.3: Story features used in all experiments.

id	Name	Type	Description
c_1	Balls	Real	Ball count similarity.
c_2	Strikes	Real	Strike count similarity.
c_3	Outs	Real	Out count similarity.
c_4	Inning	Real	Inning similarity.
c_5	Run Difference	Real	Run difference similarity.
c_6	Month	Real	Month similarity.
c_7	One Team	Binary	True if one team in common.
c_8	Two Team	Binary	True if two teams in common.
c_9	Home Run	Binary	Home run in both $\vec{g}$ and $\vec{s}$.
c_{10}	Sacrifice	Binary	Sacrifice hit in both $\vec{g}$ and $\vec{s}$.
c_{11}	Single	Binary	Single in both $\vec{g}$ and $\vec{s}$.
c_{12}	Double	Binary	Double in both $\vec{g}$ and $\vec{s}$.
c_{13}	Triple	Binary	Triple in both $\vec{g}$ and $\vec{s}$.
c_{14}	Double play	Binary	Double play in both $\vec{g}$ and $\vec{s}$.
c_{15}	Strikeout	Binary	Strikeout in both $\vec{g}$ and $\vec{s}$.
c_{16}	Fly Out	Binary	Fly out in both $\vec{g}$ and $\vec{s}$.
c_{17}	Pop Out	Binary	Pop out in both $\vec{g}$ and $\vec{s}$.
c_{18}	Ground Out	Binary	Ground out in both $\vec{g}$ and $\vec{s}$.
c_{19}	Walk	Binary	Walk in both $\vec{g}$ and $\vec{s}$.
c_{20}	Intentional Walk	Binary	Intentional walk in both $\vec{g}$ and $\vec{s}$.
c_{21}	Hit By Pitch	Binary	Hit batsman in both $\vec{g}$ and $\vec{s}$.
c_{22}	Substitution	Binary	Substitution in both $\vec{g}$ and $\vec{s}$.
c_{23}	Runner on 1st	Binary	Runner on 1st in both $\vec{g}$ and $\vec{s}$.
c_{24}	Runner on 2nd	Binary	Runner on 2nd in both $\vec{g}$ and $\vec{s}$.
c_{25}	Runner on 3rd	Binary	Runner on 3rd in both $\vec{g}$ and $\vec{s}$.
c_{26}	Marquee Matchup	Real	Confidence in this category match.
c_{27}	Great Statistics for Batter or Pitcher	Real	Confidence in this category match.
c_{28}	Bad Statistics for Batter	Real	Confidence in this category match.
c_{29}	Bad Statistics for Pitcher	Real	Confidence in this category match.
c_{30}	Opening of Inning	Real	Confidence in this category match.
c_{31}	Important Games from History	Real	Confidence in this category match.
c_{32}	Big Finish	Real	Confidence in this category match.
c_{33}	Blow Out	Real	Confidence in this category match.
c_{34}	HR hitter in 1-run game	Real	Confidence in this category match.

Table 5.4: Similarity features for $\vec{g}$ and $\vec{s}$ used in all experiments.

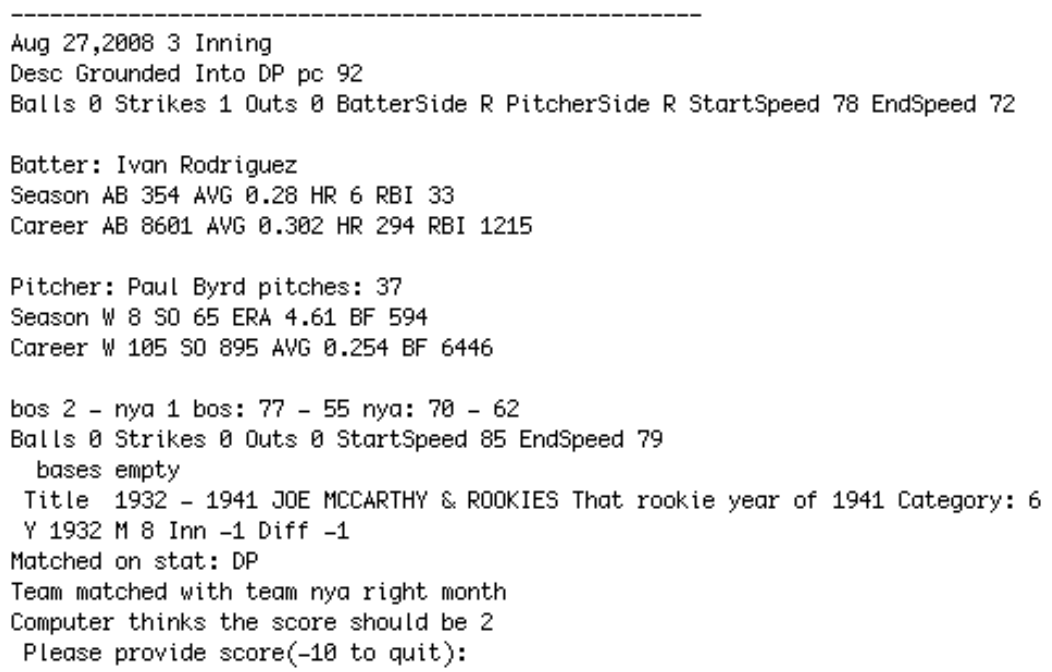

Figure 5.1: Screenshot from a program acting as an assistant to a human labelling game state and story pairs.

For IR scoring metrics, M, we considered the following: NDCG, AP, WTA, ERR, and RS. We tested the cross product of this set of scoring metrics, $k = [1, 3]$ and $y = [0, 25]$. We also tested the cross product of the full set of scoring metrics, $k = [4, 5]$ and $y = [0, 4]$, and the cross product of the full set of scoring metrics, $k = [6, 7]$ and $y = [0, 1]$.

The best performing instantiation of SCoReS had a Decision Tree as the *Evaluator*. The best set of parameters for AdaRank found using this process was $\langle M = \text{NDCG}, k = 7, y = 0\rangle$. To choose the actual *Ranker* for SCoReS we provided all 4400 training data $\mathbf{T}$ to SCoReS AdaRank. This produced the *Ranker* shown in Table 5.5. The *Contextualizer* did not allow stories to be told if there were two strikes on a batter with two outs in an inning, to avoid starting a story just before the inning ended on a strikeout. Stories also could not be told if the teams involved in the game were separated by 1 run or fewer, as there is less need for stories in close games.

This combination of *Ranker/Evaluator/Contextualizer* output stories that averaged a 3.35 match quality. Stories were only output for 23 of the 40 folds, because in 17 folds, the *Evaluator* (Decision Tree) deemed the story chosen by SCoReS AdaRank to be of insufficient quality to output. As a comparison, a perfect selector (based on the ground truth labelings) would output stories with a 3.7 average match quality, outputting a story for all

Main Sorter	α
One Team (c_7)	0.22
Run Difference (c_5)	0.14
Single (c_{11})	0.14
Out Count (c_3)	0.13
Hit by Pitch (c_{21})	0.12
Marquee Matchup (c_{26})	0.12
Great Statistics (c_{27})	0.12

Table 5.5: The *Ranker* produced by SCoReS AdaRank with 4400 training data.

40 folds. The Decision Tree operating on its own output stories that averaged a 2.3 match quality, outputting a story for all 40 folds. Thus it does not provide as good a ranking as the *Ranker*, but does provide a scalar value within the hybrid approach that can be used as a threshold to determine which top ranked stories are output by SCoReS.

While these experiments show that SCoReS performs well with respect to outputting a story with a high match quality, what we are actually interested in is the enjoyment of viewers watching games containing commentary augmented by SCoReS, as well as how useful professional commentators think SCoReS would be to them while they are commentating on games. We thus made the transition from cross-validation experiments to user studies and interviews.

5.2 User Studies

The true measure of success for SCoReS is how viewers perceive the stories it selected in the context of the game. Between summer 2011 and winter 2012, we conducted six user studies to test whether commentary adds to a broadcast, whether inserting stories into a broadcast makes it more enjoyable, and whether the added stories need to be in the proper context to add to the broadcast. Each user study involved participants watching video clips from two AAA (minor league) baseball games: the July 15, 2009 AAA All-Star game between the International League and the Pacific Coast League, and the April 7, 2011 game between the Buffalo Bisons and Syracuse Chiefs. Each study involved three different types of commentary, depending upon which hypothesis we were testing. Each video clip was between three and six minutes in length, and they were always shown in chronological order. The order of the commentary, however, varied. After each clip, participants answered questions related to their enjoyment of the clip. At the end of the session, participants completed a background questionnaire.

User Study I Questionnaire
1) I found this viewing experience enjoyable.
2) I learned something watching this clip.
3) I found this viewing experience interesting.
4) I found the video clip easy to follow.
5) I enjoyed the commentary in this clip.
6) Viewing this clip made me more interested in watching baseball.
7) Viewing this clip made me more interested in watching sports.
8) Viewing this clip made me more interested in participating in sports.

Table 5.6: User study questions asked to participants in User Study 1. Participants were asked to rate each question from 1 to 7 (strongly disagree - strongly agree).

5.2.1 User Study I – The Need for Commentary

In the first user study, we compared *SCoReS Commentary* to two different types of commentary. For *No Commentary*, we removed the commentary from the broadcast, and left the crowd noise[§]. The *Original Commentary* had voiceovers for the original professional commentary, with no stories inserted or present in the original commentary.[¶] The *SCoReS Commentary* had a story selected by our system.

We recruited 16 participants from the local community. To measure the performance of the different types of commentary, we evaluated participants' answers to the eight questions listed in Table 5.6. For each of the two games, each participant saw one clip each with *Original Commentary*, *SCoReS Commentary*, and *No Commentary*. Thus, each participant saw six video clips in total. For this experiment, SCoReS was trained on 35 game states and 88 stories, as we did not yet have our full training data. SCoReS output stories that averaged a 1.89 match quality, for 19 of the 35 cross-validation folds. As a comparison, a perfect selector would output stories with a 2.6 average match quality, outputting a story for all 35 folds. The Decision Tree (used as the *Evaluator*) output stories that averaged a 1.47 match quality over the 40 games, when used as an autonomous ranker.

The parameters for this experiment are shown in Table 5.7, with the *Ranker* being that shown in Table 5.8. Jim Prime, who did the play-by-play for this study, is an author of several baseball books, including *Ted Williams' Hit List*, a book he co-authored with Ted Williams of the Boston Red Sox [Williams and Prime, 1996]. SCoReS had a database of

[§]The crowd noise was artificial as we could not remove commentary without removing the crowd noise. After removing all the sound from the original broadcast, we added an audio file of crowd noise. This held for all types of commentary.

[¶]Voicing over the commentary was necessary as we needed to insert stories in some commentary, and needed the stories read in the same voice as the other commentary. We used voiceovers in all video clips for consistency.

Parameter	Value
Total Number of Participants	16
Baseball Fan Participants	4
Story Database Size	88
Ranker Parameters	$\langle \text{NDCG}, 4, 4 \rangle$
Play-by-Play Commentator	Jim Prime
Colour Commentator	Greg Lee

Table 5.7: Parameters for User Study I.

88 stories from which to choose stories for each game state. Figure 5.2 shows that *SCoReS Commentary* ranked significantly higher than *No Commentary* across all metrics. A one-tailed test was used to check for significance in the results of all experiments. Correcting for multiple comparisons with a Holm-Sidak test, *SCoReS Commentary* is ranked significantly higher than *No Commentary* across all metrics except "Viewing this clip made me more interested in participating in sports.". The full table of p values is shown in Appendix A.1.

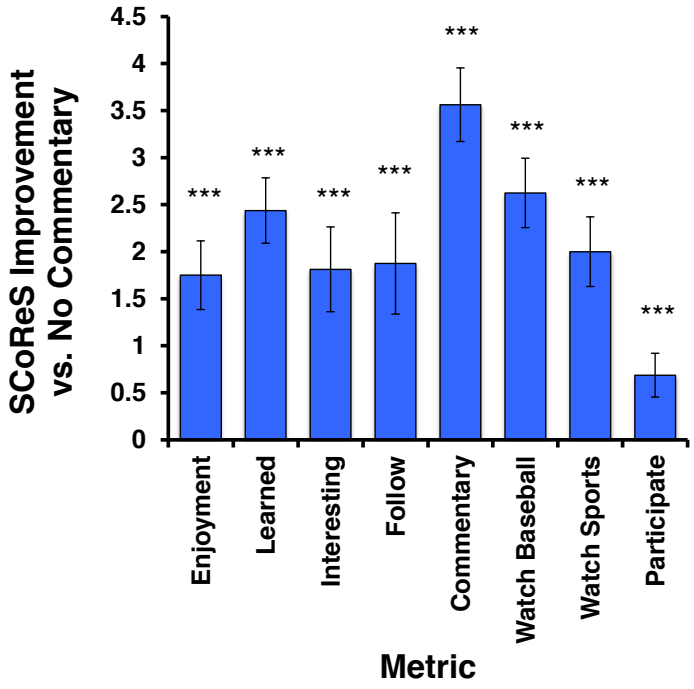

Figure 5.2: Mean (+/- Standard Error of the Mean) difference between *SCoReS Commentary* and No Commentary. *** indicates $p < 0.001$.

Main Sorter	**1st Tiebreaker**	**2nd Tiebreaker**	**3rd Tiebreaker**	α
One Team (c_7)	Two Teams (c_8)	Triple (c_{13})	Walk (c_{19})	0.36
Run Diff (c_5)	One Team (c_7)	Strikeout (c_{15})	Marquee(c_{26})	0.24
Two Teams (c_8)	Hit by Pitch (c_{21})	Runner on First (c_{23})	Strikeout (c_{15})	0.21
Out (c_3)	One Team (c_7)	Home Run (c_9)	Int Walk (c_{20})	0.19

Table 5.8: The *Ranker* produced by SCoReS AdaRank with 3080 data.

5.2.2 User Studies II, III, & IV – Discovery of the First Clip Bias

Having collected evidence that commentary itself adds entertainment to a game broadcast, we performed two more user studies, replacing *No Commentary* with *Mismatch Commentary*. For *Mismatch Commentary*, we inserted a story in the same place as we would with SCoReS, but instead of the story being selected for that game state, it was actually a story chosen by SCoReS for a different game state, in the other game the participant saw. This allowed us to keep the story pool consistent across the conditions thereby controlling for the overall level of interest.

User Studies II and III

User Studies II and III took place concurrently. User Study II was conducted at the University of Alberta (U of A), while User Study III took place at Acadia University. Due to the procedural differences in using the subject pool at the two institutions, participants from the U of A did not know the subject matter of the study beforehand, while those at Acadia University did (thus, they were "self-selected"). Both sets of participants were drawn from the Psychology subject pool at their respective academic institutions. To better assess the value of the different types of commentary, the questions asked of participants were updated for User Studies II and III (Table 5.9) [‡]. Questions from Table 5.6 deemed unnecessary were removed, and questions pertaining specifically to the story were added in an attempt to gauge what effect the stories were having on viewers' enjoyment of the clips. As in the first user study, participants in User Studies II and III saw six video clips each, three from each game and two of each type of commentary.

Parameters for User Studies II and III (as well as IV) are given in Table 5.11, with the *Ranker* given in Table 5.12. The cross-validation process described in section 5.1 was updated to perform 10 iterations of each parameter set and take the average, to avoid bias from random tie-breaking in rankings (in the case where the ranker itself does not provide

[‡]There were also four "dummy" questions asked in an attempt to prevent subjects from deducing exactly what we were testing. These are shown in Appendix D.6.

User Studies II, III, V, and VI Questionnaire
1) I found this viewing experience enjoyable.
2) I learned something about baseball history watching this clip.
3) I found the video clip easy to follow.
4) I enjoyed the commentary in this clip
5) Viewing this clip made me more interested in watching baseball.
6) I thought the story in this clip was enjoyable.
7) I thought the story in this clip was fitting to the game at hand.
8) I thought the story in this clip was annoying.

Table 5.9: Questions asked to participants in User Studies II, III and IV.

enough information to completely order the stories). This implementation of SCoReS averaged a 2.25 match quality over 16 game states for which it output a story, while (as in User Study I) a perfect ranker would have averaged a 2.6 match quality over all 35 folds. In an attempt to improve the *Ranker* and *Evaluator* performance, we introduced 18 new similarity features. None of the *Rankers* used in he user studies made use of these features but the *Evaluator* (a Decision Tree) did use these features in some cases. These similarity features are listed in Table 5.10.

In an attempt to overcome any ordering biases, nine different combinations of the commentary orderings were used in the study (Table 5.13). By having approximately the same number of participants see each of these combinations, we ensure that each type of commentary appears in each chronological position the same number of times for each game. Also, each SCoReS selected story from game A appears as a Mismatch story in game B the same number of times (3) and at each place in the ordering the same number of times (1). As an example the SCoReS chosen story for the second clip of Buffalo Bisons vs Syracuse Chiefs game (labelled $\mathcal{M}_2$ when shown in the All-Star game) appears first in Group 2, second in Group 5 and third in Group 9. It is also important to ensure that no subject hears the same story twice. Thus, $\mathcal{M}_2$ cannot be shown in the All-Star game in any group where *SCoReS Commentary* appears second for the Buffalo Bisons - Syracuse Chiefs game (Groups 1, 3 and 6). While the order of the games is less important, we did play Game A first in odd groups and Game B first in even groups. The results are discussed below.

User Study IV

User Study IV differed from the first three studies in several aspects. Firstly, participants watched three video clips from one game, and then were given a choice concerning which version of the fourth clip they would like to watch. They were told that the three clips

id	Name	Type	Description
c_{36}	Career Average Against Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ career BAA.
c_{37}	Season Average Against Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ season BAA.
c_{38}	Career Wins Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ career W.
c_{39}	Season Wins Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ season W.
c_{40}	Career Batting Average Measure	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ career BA.
c_{41}	Season Batting Average Measure	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ season BA.
c_{42}	Career Home Runs Measure	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ career HR.
c_{43}	Season Home Runs Measuret	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ season HR..
c_{44}	Last Career Average Against Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ last career BAA.
c_{45}	Last Season Average Against Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ last season BAA.
c_{46}	Last Career Wins Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ last career W.
c_{47}	Last Season Wins Measure	Real	($\vec{s}$ K+NH) $\times$ $\vec{g}$ last season W.
c_{48}	Last Career Batting Average Measure	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ last career BA.
c_{49}	Last Season Batting Average Measure	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ last season BA.
c_{50}	Last Career Home Runs Measure	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ last career HR.
c_{51}	Last Season Home Runs Measuret	Real	($\vec{s}$ HR+2B+3B+GS) $\times$ $\vec{g}$ last season HR..
c_{52}	Season Batting Average Relevance	Real	$\#$ of untrue $\vec{s}$ features $\times$ $\vec{g}$ season BA.
c_{53}	Season Average Against Relevance	Real	$\#$ of untrue $\vec{s}$ features $\times$ $\vec{g}$ season BAA.

Table 5.10: Similarity features added after User Study I. Here "K" = "strikeout", "NH" = "no-hitter", "BA" = "batting average", "BAA" = "batting average against", "HR" = "home run", "2B" = "double", "3B" = "triple", "GS" = "grand slam" , "W" = "wins" and "Last" = "Pitcher/Batter for the previous pitch".

Parameter	User Study II	User Study III	User Study IV
Institution	U of A	Acadia	Acadia
Total Number of Participants	97	28	22
Baseball Fan Participants	17	21	13
Story Database Size	88	88	88
Ranker Parameters	$\langle$AP, 3, 2$\rangle$	$\langle$AP, 3, 2$\rangle$	$\langle$AP, 3, 2$\rangle$
Play-by-Play Commentator	Greg Lee	Greg Lee	Len Hawley
Colour Commentator	Jim Prime	Jim Prime	Jim Prime

Table 5.11: Parameters for User Studies II, III and IV.

Main Sorter	Tiebreaker	α
One Team (c_7)	Big Finish (c_{32})	0.57
Sacrifice (c_{10})	Bad Pitcher Confidence (c_{28})	0.22
Blowout (c_{33})	Home Run Hitter One Run Game (c_{34})	0.21

Table 5.12: The *Ranker* produced by SCoReS AdaRank for User Studies II, III, and IV.

Group	All-Star Audio Ordering	Bisons Audio Ordering	First Game
Group 1	$\mathcal{O}\ \mathcal{M}_1\ \mathcal{S}$	$\mathcal{M}_2\ \mathcal{S}\ \mathcal{O}$	A
Group 2	$\mathcal{M}_2\ \mathcal{O}\ \mathcal{S}$	$\mathcal{S}\ \mathcal{M}_2\ \mathcal{O}$	B
Group 3	$\mathcal{O}\ \mathcal{S}\ \mathcal{M}_3$	$\mathcal{M}_1\ \mathcal{S}\ \mathcal{O}$	A
Group 4	$\mathcal{O}\ \mathcal{S}\ \mathcal{M}_1$	$\mathcal{M}_3\ \mathcal{O}\ \mathcal{S}$	B
Group 5	$\mathcal{S}\ \mathcal{M}_2\ \mathcal{O}$	$\mathcal{S}\ \mathcal{O}\ \mathcal{M}_3$	A
Group 6	$\mathcal{M}_3\ \mathcal{S}\ \mathcal{O}$	$\mathcal{O}\ \mathcal{S}\ \mathcal{M}_1$	B
Group 7	$\mathcal{S}\ \mathcal{M}_3\ \mathcal{O}$	$\mathcal{S}\ \mathcal{O}\ \mathcal{M}_2$	A
Group 8	$\mathcal{M}_1\ \mathcal{O}\ \mathcal{S}$	$\mathcal{O}\ \mathcal{M}_1$ R	B
Group 9	$\mathcal{S}\ \mathcal{O}\ \mathcal{M}_2$	$\mathcal{O}\ \mathcal{M}_3\ \mathcal{S}$	A

Table 5.13: User Study II and III Orderings. *Original Commentary* = $\mathcal{O}$, *SCoReS Commentary* = $\mathcal{S}$ and *Mismatch Commentary* = $\mathcal{M}$.

they had seen featured different types of commentary and asked which of these types of commentary they would like to see in the fourth clip. They were not explicitly told which types of commentary were available, simply that they could choose between the types of commentary in the first three clips they watched. They then watched the fourth clip and answered whether they were pleased with their choice, or not. User Study IV took place at Acadia, again with self-selected participants. Len Hawley, the play-by-play commentator for the Acadia Men's Varsity hockey team did the play-by-play reading for this experiment, and the last two user studies as well. This provided the experiments with a professional voice, and removed any bias created by having the researcher do some of the commentating.

First Clip Bias

The ordering of the commentary types in User Study IV (*Original Commentary, SCoReS Commentary and Mismatch Commentary*) was balanced, so that each type appeared in each clip approximately the same number of times. The type of commentary from the first clip shown was never chosen for the fourth clip by a participant, however. A similar issue was noted in User Studies II and III – participants rated the first clip they saw statistically significantly lower for all questions, except "I learned something about baseball history watching this clip". *SCoReS Commentary* did score significantly higher than *Original Commentary* and *Mismatch Commentary* for this metric, with $p < 0.05$. Beyond this metric though, results for studies II – IV were disregarded, as the *first clip bias* distorted participants perceptions of each commentary type. Data supporting the first clip bias can be found in Appendix A.2, while some discussion of results for User Studies II and III can be found in Appendices A.3 and A.4, respectively. The first clip bias was not seen in User Study I,

which is why those results were presented.

5.2.3 User Studies V and VI – Baseball Fans prefer SCoReS

Similarly to User Studies II and III, User Studies V and VI took place concurrently at the University of Alberta and Acadia University, respectively. Participants from the U of A again had no knowledge of study subject matter beforehand, and Acadia students again were self-selected.

The setup for User Studies V and VI included some adjustments to previous studies. Foremost amongst these was eliminating the first clip bias. To do so, we inserted a "dummy clip", which preceded the three video clips of interest. Thus, each participant saw eight video clips in total. Another adjustment was to screen participants with the question "Are you a baseball fan?" [†]. As baseball fans actually enjoy baseball games, we hypothesized that they would be likely to notice differences in commentary. Also, SCoReS is built to improve the enjoyment of a sport. If someone does not enjoy a sport to begin with, it may be difficult to change his or her mind. The ordering of video clips was again that shown in Table 5.13. SCoReS had access to the full 4400 training data, and output the *Ranker* shown in Table 5.5. Parameters for these experiments are given in Table 5.14.

We recruited 39 students from the U of A and 17 students from Acadia University who were self-described baseball fans. For each of the two games, each participant saw one clip each with *Original Commentary*, *SCoReS Commentary*, and *Mismatch Commentary*. The 17 baseball fans from Acadia did not yield any significant results, likely mostly due to the small number of them. The baseball fans from the University of Alberta, did, however. Figure 5.3 shows the mean difference between *SCoReS Commentary* and both *Original Commentary* and *Mismatch Commentary* for User Study VI. *SCoReS Commentary* was ranked higher than the *Original Commentary* for the "Viewing this clip made me more interested in watching baseball" metric, with $p < 0.001$. This shows that adding stories to commentary can improve a broadcast. *SCoReS Commentary* was ranked higher than *Mismatch Commentary* for the "I found this viewing experience enjoyable" metric with $p < 0.01$. This shows that intelligently adding stories to commentary can be more enjoyable to the viewer than adding random stories.

Correcting for multiple comparisons with a Holm-Sidak test, $p < 0.05$ for both of these comparisons. The full list of p values can be found in Appendix A.5. Questions 6

[†]We asked several screening questions and allowed subjects who said "yes" to any of them to participate, hence why not all the participants in these studies were baseball fans.

Parameter	User Study V	User Study VI
Institution	Acadia	U of A
Total Number of Participants	23	69
Baseball Fan Participants	12	39
Story Database Size	110	110
Ranker Parameters	$\langle \text{WTA}, 7, 0 \rangle$	$\langle \text{WTA}, 7, 0 \rangle$
Play-by-Play Commentator	Len Hawley	Len Hawley
Colour Commentator	Jim Prime	Jim Prime

Table 5.14: Parameters for User Studies V and VI.

– 8 from Table 5.9 are omitted from the graph as 1) the questions were irrelevant to the *Original Commentary* case and b) there were no significant differences between *SCoReS Commentary* and *Mismatch Commentary* for these questions.

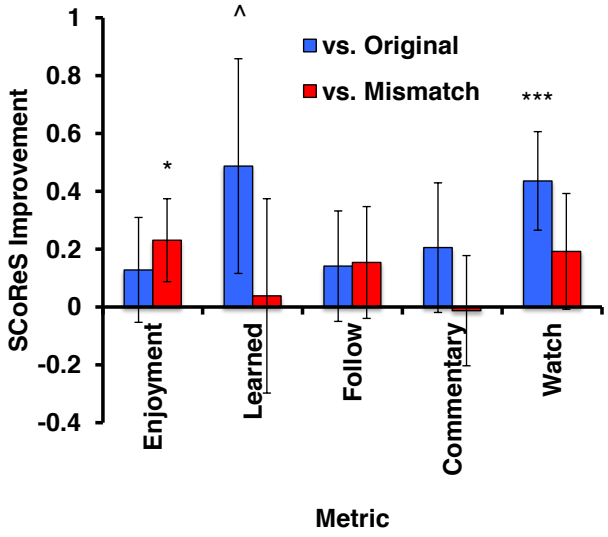

Figure 5.3: Mean (+/- Standard Error of the Mean) difference between *SCoReS Commentary* and *Original Commentary* or *Mismatch Commentary*. *** indicates $p < 0.001$; * indicates $p < 0.05$; $\wedge$ indicates $p < 0.1$.

5.3 Interviews with Commentators

In this experiment, we demonstrated SCoReS to professional commentators. To evaluate the potential usefulness of SCoReS, we first asked them, "Would you be interested in a system that suggests interesting stories during a game?". Then we demonstrated SCoReS delivering three stories for four different clips to the commentators. After each clip, we asked, "Would you tell any of the suggested stories?". The commentators could answer based on a synop-

sis, or choose to see the full story text. After the full demonstration we asked, "Would you be interested in this particular system?". The *Ranker* used within SCoReS to choose stories for the demonstration was the same as that used in User Studies V and VI (Table 5.5). A screenshot of the demonstration is shown in Figure 5.4.

The four commentators were Len Hawley (who did play-by-play reading for some video clips, but was detached from the story insertion aspect of the user studies), Dan Robertson, play-by-play commentator for various sports (including baseball) for Eastlink Television, and Mark Lee and Kevin Weekes, a play-by-play an colour commentator team for the Canadian Broadcasting Corporations's (CBC's) *Hockey Night in Canada* programme.

All four commentators said they believed a system such as SCoReS would be a useful tool to have at their disposal. When asked about SCoReS itself, they all answered that it would be a great tool not only for baseball, but also for other sports, with a few tweaks. In particular, stories would need to be kept short for hockey broadcasts, which is a faster moving sport.

Among the four clips shown to each commentator, a story suggested in two of them would have been told by Len Hawley and Dan Robertson with some minor changes. One of these was a story about a player having a successful day batting, the day after his divorce in late April, 1988 (the full story can be found in Appendix B.1). The game state where the story was told was during the April 7th Buffalo Bisons game, just after a player had his fourth hit and fifth run batted in of the game, thus, a successful game in his own right. Both commentators said they would have told the story if it had happened in early April. The *Ranker* within SCoReS did not make use of the month match feature, c_6, for this experiment. Thus, the story happening earlier in April would have had no effect on SCoReS, and this story could still have been output. Even if c_6 were part of the *Ranker* used, c_6 only takes account of the month of a story, not the day, so SCoReS would again have been unaffected by this new knowledge.

Mark Lee and Kevin Weeks also said in two of the clips shown that they would have told one of the stories presented by SCoReS, with no conditions on their answers. They offered several other insights as well. First, even if the stories suggested are not appropriate at the time of suggestion, once commentators have seen them, they can keep the stories in mind in case they are relevant later. As Mark Lee pointed out, on the two occasions he did not choose to tell a story, SCoReS suggested stories with a borderline match quality to the current game state, but if the game state were to change slightly, the match quality would be high enough for the story to warrant telling. Secondly, a system such as SCoReS would

Figure 5.4: Screenshot of stories suggested by SCoReS overlaid on the screen during the 2009 AAA All-Star game.

need to be completely integrated into a broadcast, and not simply used by the commentating team. This would allow the producers to add imagery to the broadcast, relevant to the story that the commentators choose to tell. Thirdly, while commentators do know many stories about the sport being broadcast, there are too many different parts of the game they must monitor and process in order to properly broadcast the game that they often cannot think up stories on their own. Thus SCoReS can be beneficial to the broadcasters not only in terms of suggesting stories they do not know, but also in terms of stories they *do* know, but would not have thought of and connected to the current game state. Lastly, while older stories may seem less relevant to younger viewers, they are actually quite relevant to viewers who watched games at the time the stories took place. Thus, SCoReS can help keep the interest of several generations of sports fans by connecting a sport's past to its present.

Chapter 6

Discussion and Conclusion

We have shown that SCoReS has a statistically significant positive influence on the sports viewing experience across several metrics. National commentators Mark Lee and Kevin Weekes were particularly positive about the system, suggesting its appeal for a broad audience. This indicates that implementing SCoReS in professional commentating may lead to a better viewing experience.

6.1 Lessons Learned

Participants in the user studies were recruited from subject pools, and study participation made up a part of their grade. Thus, they did not watch the video clips at a time of their choosing, as opposed to a fan watching on television. A more effective way of evaluating SCoReS would be to have baseball fans evaluate the system at a time of their choosing, when they are interested in watching baseball.

Viewing three to six minute clips of games can make it difficult for a participant to gain context in the game, and thus make it difficult to appreciate an appropriately placed story. An ideal setup for SCoReS evaluation would be to have participants watch an entire baseball game, so that they could better gauge whether a particular story should be told at a particular time.

Professional commentators generally state why stories they are telling are relevant, to give context to the story. This did not happen during the user studies, because we believed this would have biased participants' answers to some of the questions. In hindsight, it may have been possible to state why a story was being told by mentioning the game features that led to the story's selection, for both the SCoReS stories and the Mismatch stories.

Despite these challenges, SCoReS was able to achieve significant improvements in overall enjoyment and increasing interest in watching baseball, and we surmise that in

a more realistic deployment, SCoReS would further improve the entertainment value of sports broadcasts. Identifying these challenges and the ways to overcome them is one of the contributions of this project and we hope these will be used by future researchers.

6.2 Future Research

In the future, we would like to augment SCoReS with baseball facts, in addition to stories. These facts would be shorter bits of information from baseball's past, that could be told quickly. Also, the system would benefit from an automated bot to perform several tasks. The bot could mine the web looking for relevant stories, providing SCoReS with an ever-growing set of stories from which to choose. The bot could also automatically extract features from the stories it finds, eliminating the need to do so by hand.

Adding more training data (both game states and stories) should also benefit SCoReS, by helping to create a more intelligent learner. A larger story database for SCoReS to choose from online should provide for a system with more flexibility, making it more likely there is a story in the database relevant to each game situation. Data labelling is a time-consuming process that could be performed by multiple individuals (rather than just one domain expert), possibly making use of such services as Mechanical Turk [Amazon, 2012].

6.3 Future Applications

SCoReS offers many possible future applications along the lines of fully automated commentary. Combining SCoReS with the systems described in Chapter 3 (such as *Byrne*) would yield a completely automated commentating system. Sports video games could increase their story databases, and then use SCoReS to select between these stories during gameplay. Automated storytelling systems such as Statsheet and Narrative Science could use SCoReS to automatically add stories to their own automatic recaps of games.

SCoReS could also be used to created personalized colour commentary. The viewing experience could be tailored to not just simply groups, but individuals, through web broadcasts. Features describing the viewer could be input into the system, and stories could be selected partly based on these features. Finally, SCoReS could be used for story selection in real-time strategy games such as *StarCraft* as they have their own sets of features and stories, and often contain commentary.

6.4 Conclusion

Storytelling is believed to be a cognitively rich and creative task. In order to excel in storytelling, an innate aptitude and training are required. Skilled storytellers including writers, poets and colour commentators are recognized and famed. In this work, we took a step towards automating this task by building the first AI story selector for colour commentary in any sport. Its implementation in baseball was positively evaluated in user studies and by national level professional commentators. We believe this to be a contribution to the field of Artificial Intelligence with immediate practical applications.

Appendix A

Supplemental Results for User Studies

In this appendix, we provide p values for User Studies I and VI, and supplemental results for user studies affected by the first clip bias. These results helped guide us in later experiments (User Studies V and VI, and commentator interviews) and also provide some evidence that SCoReS is able to improve the quality of a broadcast.

Metric	p **value**	**Corrected** p **value**
Enjoyment	1.6×10^{-7}	1.3×10^{-6}
Learned	1.1×10^{-6}	8.8×10^{-6}
Follow	8.5×10^{-7}	6.8×10^{-6}
Commentary	9.1×10^{-5}	7.3×10^{-4}
Watch Baseball	2.0×10^{-8}	1.6×10^{-7}
Interesting	1.4×10^{-7}	1.1×10^{-6}
Watch Sports	2.0×10^{-6}	1.6×10^{-5}
Participate	8.5×10^{-3}	6.6×10^{-2}

Table A.1: p values for the bars in Figure 5.2. All differences are significant ($p < 0.05$), even with the Holm-Sidak correction, except for the "Participate" metric.

A.1 User Study I

Table A.1 shows p values obtained from a one-tailed t-test on the data from User Study I.

A.2 First Clip Bias

Figure A.1 shows the effect of the first clip bias (see Section 5.2.2) on the Enjoyment metric in User Study II. The first clip is ranked significantly lower, regardless of commentary type. The rest of the metrics are ranked similarly, except for the "I learned something about

baseball history watching this clip". This bias is likely due to participants being unfamiliar with the teams and game when the watch the first clip, and becoming more familiar after having seen the first clip.

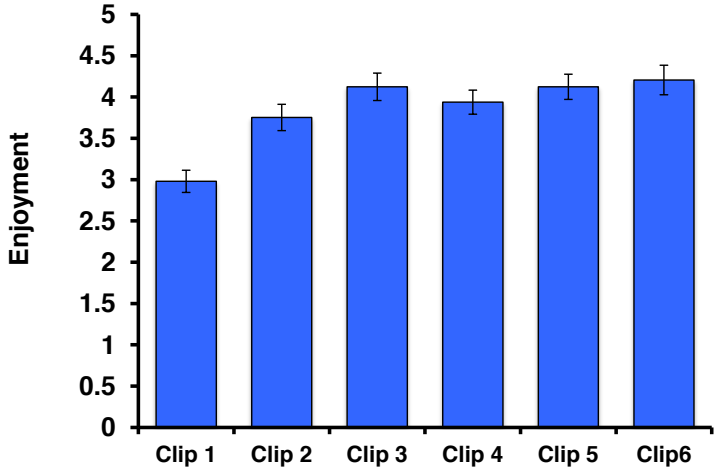

Figure A.1: The enjoyment of participants per clip, for User Study II. Clip 1 suffers from the first clip bias, as it is ranked significantly lower than Clips $2-5$.

Table A.2 shows the difference between the scores for the first clip and the scores for the clip that scored the worst (other than clip 1), for the enjoyment and commentary metrics. User Studies, II, III and IV are all significantly affected by the first clip bias, while User Study I is not. User Studies V and VI do not suffer from the first clip bias due to the showing of a dummy clip before the first clip of interest. Tables for metrics other than the "Learned" metric show similar results.

A.3 User Study II

User study II took place at the University of Alberta. Results from these experiments were affected by the first clip bias, rendering the results untrustworthy. Still, trends seen in the

Study	**Enjoyment Difference**	**Enjoyment p value**	**Commentary Difference**	**Commentary p value**
I	-0.125	4×10^{-1}	$+0.19$	4×10^{-1}
II	$+1.143$	1×10^{-6}	$+1.5$	3×10^{-6}
III	$+0.773$	1×10^{-7}	$+1.37$	8×10^{-14}
IV	$+1.000$	4×10^{-3}	$+0.8$	4×10^{-2}
V	0	5×10^{-1}	$+0.12$	3×10^{-1}
VI	-0.045	6×10^{-1}	-0.18	8×10^{-1}

Table A.2: The mean difference between the minimum of clips $2-5$ and clip 1 over several metrics.

results from these studies led us to the successful User Studies IV and V, so we present the results here.

Figure A.2 shows the mean difference between *SCoReS Commentary* and both *Original Commentary* and *Mismatch Commentary* for 97 participants at the University of Alberta. Recall that the only metric unaffected by the first clip bias was that measuring "I learned something watching this clip". As Figure A.2 shows, *SCoReS Commentary* outperforms both *Original Commentary* and *Mismatch Commentary* by a statistically significant margin in this metric. This makes intuitive sense for the *Original Commentary* clip, as it contains no story, while the clip containing *SCoReS Commentary* does. In the case of *Mismatch Commentary*, it is possible that participants learned more from the *SCoReS Commentary* case because the story they heard had some connection to the game at hand, thus keeping their attention.

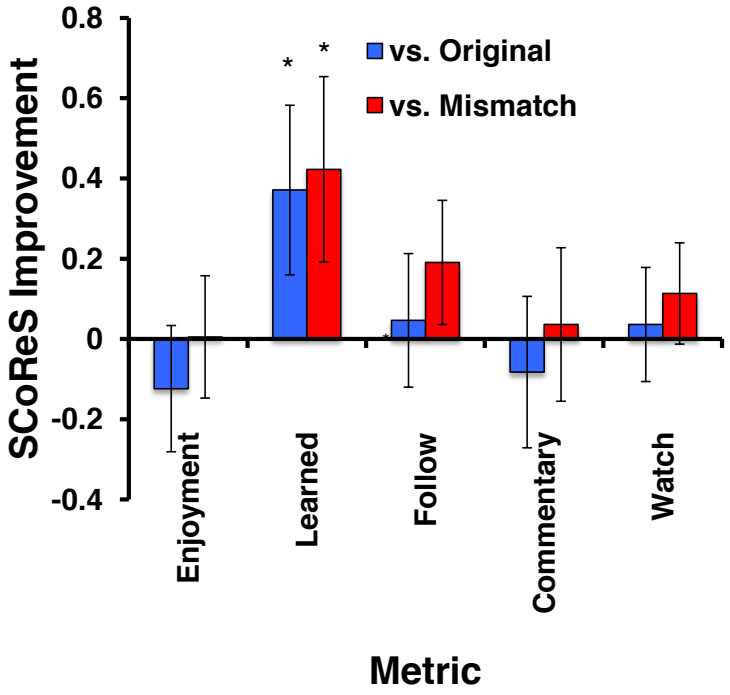

Figure A.2: Mean (+/- Standard Error of the Mean) difference between *SCoReS Commentary* and *Original Commentary* or *Mismatch Commentary*, for User Study II. * indicates $p < 0.05$.

A.4 User Study III

Figure A.3 shows the mean difference between *SCoReS Commentary* and both *Original Commentary* and *Mismatch Commentary* for 28 participants at the Acadia University. User Study III was originally meant as a supplemental study to User Study II – to boost the

Question	vs Original	vs Mismatch	Corrected Original	Corrected Mismatch
Enjoyed	1.8×10^{-1}	1.0×10^{-2}	6.2×10^{-1}	5.0×10^{-2}
Learned	6.0×10^{-2}	4.4×10^{-1}	2.7×10^{-1}	9.4×10^{-1}
Follow	1.2×10^{-1}	1.3×10^{-1}	4.9×10^{-1}	5.0×10^{-1}
Commentary	1.0×10^{-1}	5.3×10^{-1}	4.3×10^{-1}	9.8×10^{-1}
Watch	8.0×10^{-4}	1.0×10^{-1}	4.0×10^{-3}	4.2×10^{-1}

Table A.3: The p values when comparing *SCoReS Commentary* to *Original Commentary* and *Mismatch Commentary* in User Study VI (for the bars in Figure 5.3).

participant number and completely balance the study with respect to the orderings in Table 5.13. We later deemed it a separate study due to the self-selection of the participants at Acadia. This led to User Study III being unbalanced in terms of which clip was seen first. Among the 28 participants, only 2 saw the a first clip with *SCoReS Commentary*, while 13 each saw *Original Commentary* and *Mismatch Commentary* first. Given the first clip bias, SCoReS was given a tremendous advantage in this study, and this could explaine SCoReS showing improvement over the other two types of commentary, as seen in Figure A.3.

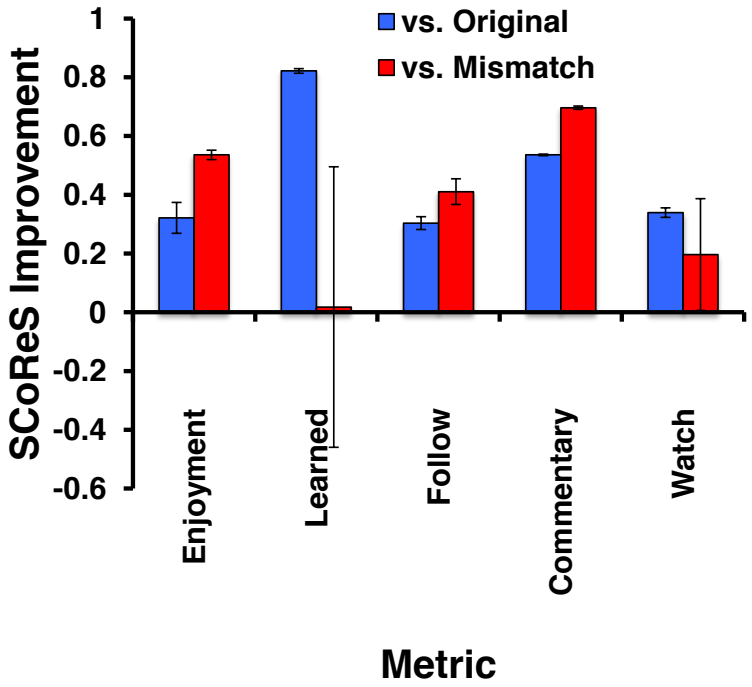

Figure A.3: Mean (+/- Standard Error of the Mean) difference between *SCoReS Commentary* and *Original Commentary* or *Mismatch Commentary* for User Study III. We omit significance symbols due to the unbalanced nature of the experiment.

A.5 User Study VI

Table A.3 shows the p values for Figure 5.3, the results for User Study VI, as well as the p values after being corrected with a Holm-Sidak multiple comparison correction.

Appendix B

Sample Stories

The following four stories were part of the story database used in all experiments. They were among the stories suggested to professional commentators as part of the demonstration of SCoReS.

B.1 Keith Hernandez's Divorce

Former New York Met Keith Hernandez likes to tell the story about the day after his divorce. Hernandez's divorce was effective on April 25, 1988, and in the Mets' 13-4 victory the next day over the Braves, Hernandez hit two homers, including a grand slam, and collected seven RBIs. After the game, he told reporters, "If I got a divorce every day, I'd be in the Hall of Fame."

B.2 1986 World Series

In the 1986 World Series, the Boston Red Sox led the New York Mets by three games to one and by two runs in the bottom of the tenth inning of Game 6. With two outs, the Mets hit three consecutive singles, and then benefitted from a Bob Stanley wild pitch to tie the game. This left them with a runner at second base and two out, for the batter, Mookie Wilson. After fouling off several pitches, Wilson hit a weak grounder down the first base line that somehow eluded first baseman Bill Buckner and led to the Mets winning the game. The Mets would win Game 7 to take the 1986 World Series.

B.3 Red Barrett's Quick Game

Red Barrett was a pitcher for the Boston Braves known for keeping the clubhouse loose with his antics. He also was known for pitching quick games. On August 10, 1944, he shut

out the Cincinnati Reds 2-0, allowing only 2 singles. The entire game took just one hour and 15 minutes to play. He threw only 58 pitches, a major league record for a complete game.

B.4 Luis Tiant's Tough All-Star Loss

Luis Tiant, who pitched for the Cleveland Indians and Boston Red Sox among other teams, had his best season in 1968, going 21-9 with 9 shutouts and a 1.60 ERA. He had some rough luck in the Major League All-Star game, though. Chosen to start the game, he gave up a lead off single to Willie Mays. Tiant then attempted a pick-off throw, but first baseman Harmon Killebrew was not prepared, and missed the throw. Mays advanced to second, then to third on a wild pitch from Tiant that walked Curt Flood. The double play that followed allowed Mays to score. This would be the only run of the game, as the National League prevailed 1-0, the first time an AllStar game ended with that score. Tiant later said "I lost the game because I tried to pick-off Mays. Instead, I picked off Killebrew".

Appendix C

Marquee Matchup Feature Calculation

In this appendix, we present how the marquee matchup feature is calculated, referencing the table of similarity features (Table 5.4) the table of game state features (Table 5.1) and the table of story features (Table 5.3).

In order to compute the marquee matchup feature c_{26}, we first calculate four intermediate features f to later combine into a single scalar value. Table C.1 shows the threshold values for f_1, the *Good Batting Average* feature, while Table C.2 shows the threshold values for f_2, the *Good Pitcher Average Against* feature.

Test on Season Batting Average (g_7)	f_1
$g_7 \leq 0.26$	0.3
$g_7 \leq 0.28$	0.5
$g_7 \leq 0.30$	0.7
$g_7 > 0.30$	1

Table C.1: Calculation of the good batting average feature f_1.

Test on Season Pitching Average (g_{12})	f_1
$g_{12} \leq 0.23$	1
$g_{12} \leq 0.25$	0.8
$g_{12} \leq 0.27$	0.5
$g_{12} \leq 0.28$	0.3
$g_{12} > 0.28$	0

Table C.2: Calculation of the good pitching average against feature f_2.

The *Good Batter Home Run Total* feature, f_3 is calculated according to the formula:

$$f_3 = g_8 \Big/ \Big(\frac{(g_{32} - 4) \cdot 30 + g_{33}}{180} \cdot 30 \Big)$$

where g_8 is the batter's season home run total, g_{32} is the current month and g_{33} is the current day. This formula considers a total of 30 home run total for a season as a marquee home run hitter, and uses the day and month features to correct for the time of season during which the game is taking place. The baseball regular season runs from April (the fourth month of the year) through September (the ninth month), and there are approximately 30 days in a month (with exactly 30 in September), giving this calculation the approximate range of $[0, 1]$. This feature assumes a 30 home run pace is optimal. A home run pace over 30 is capped at 1.

The *Good Pitcher Win Total* feature f_4 is calculated in a similar manner:

$$f_4 = g_{10} \Big/ \Big(\frac{(g_{32} - 4) \cdot 30 + g_{33}}{180} \cdot 15 \Big)$$

where g_{10} is the pitcher's win total. Like a high home run pace, a win total pace over 15 is capped at 1.

The marquee matchup feature c_{26} is first set to $0.25 \cdot (f_1 + f_2 + f_3 + f_4)$, considering each intermediate feature f equal. Thus, c_{26} has range $[0, 1]$. As the value of c_{26} is not simply a representation of whether the game state presents a marquee matchup, but a confidence in a marquee matchup match, its value is further refined by the category of the story, as shown in Table C.3.

Category feature s_{45} value	**Category**	**Refinement of Marquee Matchup Feature c_{26}**
1	Opening of Inning	$c_{26} \leftarrow c_{26} \cdot 0.5$
2	Bad Hitter	$c_{26} \leftarrow c_{26} \cdot 0.1$
3	Marquee Matchup	$c_{26} \leftarrow c_{26} \cdot 1.0$
4	Great Stats	$c_{26} \leftarrow c_{26} \cdot 0.8$
5	Bad Stats	$c_{26} \leftarrow c_{26} \cdot 0.1$
6	Important Game	$c_{26} \leftarrow c_{26} \cdot 0.3$
7	Big Finish	$c_{26} \leftarrow c_{26} \cdot 0.3$
8	Blowout	$c_{26} \leftarrow c_{26} \cdot 0.3$
9	Home Run Hitter	$c_{26} \leftarrow c_{26} \cdot 0.3$
10	Human Interest	$c_{26} \leftarrow c_{26} \cdot 1.0$

Table C.3: Refinement of the marquee matchup feature c_{26} according to the category of the story s_{45}.

If the story category has little in common with a marquee matchup (e.g., "Bad Hitter"), then the marquee matchup confidence is diminished. The more similar a category is to

a marquee matchup, the less its value is diminished. The exception is “Human Interest”, which is a catch-all category of sorts. No category confidence is diminished for when this is the story category. The other category confidence features (c_{27} – c_{34}) are calculated in a similar manner.

Appendix D

User Study Materials

D.1 Briefing for User Studies I,II and III

Introduction: Welcome! You are invited to participate in a research study being conducted by Greg Lee and Dr. Vadim Bulitko of the Department of Computing Science from the University of Alberta. The purpose of this study is to evaluate the effectiveness of different types of sports commentary.

Your participation: Your participation in this study involves watching six 5-minute video clips of Minor League Baseball games. Following the completion of the clip, you will be asked to fill out a survey ranking the games across several measures, including which game was more enjoyable to watch.

Your rights: Your decision to participate in this study is entirely voluntary and you may decide at any time to withdraw from the study. Your decision to discontinue will not affect your academic status or access to services from the University of Alberta. If you choose to participate, you may skip any items you do not wish to answer. Responses made by individual participants on the questionnaires will remain confidential, and your name will not appear on the questionnaire or be associated with your responses in any way. Questionnaires will be identified only by a researcher-assigned code number, for the purpose of associating survey forms with the particular story that the participant experienced. Only researchers associated with the project will have access to the questionnaires. The results of this study may be presented at scholarly conferences, published in professional journals, or presented in class lectures. All data presented will be anonymous. The data will be securely stored by Greg Lee) for a minimum of five years.

Benefits and risks: There are no foreseeable risks to this study, but if any risks should arise, the researcher will inform the participants immediately. If you should experience any adverse effects, please contact Greg Lee and/or Dr. Vadim Bulitko immediately.

Contact information If you have any questions or comments on the study, or if you wish a clarification of rights as a research participant, you can contact Greg Lee or the Human Research Ethics Committee at the number and address below.

Greg Lee
Ph.D. Candidate
27 Finch Court
Kentville, NS
(902) 365-5884

Vadim Bulitko, Ph.D.
Associate Professor
Department of Computing Science
University of Alberta Edmonton, AB
T6G 2E8
(780) 492-3854

University of Alberta Research Ethics Office
(780) 492-2614
reoffice@ualberta.ca

Stephen Maitzen, Ph.D.
Chair of the Acadia Research Ethics Board
Department of Philosophy, Acadia University
Wolfville, NS
B4P 2R6
(902) 585-1407

Signatures. Please sign below to indicate that you have read and understood the nature and purpose of the study. Your signature acknowledges the receipt of a copy of the consent form as well as indicates your willingness to participate in this study.

D.2 Briefing for User Studies IV,V and VI

Introduction: Welcome! You are invited to participate in a research study being conducted by Greg Lee and Dr. Vadim Bulitko of the Department of Computing Science from the University of Alberta. The purpose of this study is to evaluate the effectiveness of different types of sports commentary.

Your participation: Your participation in this study involves watching four 5-minute video clips of Minor League Baseball games. Following the completion of each clip, you will be asked to fill out a survey ranking the clip across several measures, including how enjoyable it was to watch. In addition, following the first three clips you will be asked to choose which version of the fourth clip you would like to see, based on what you saw in the first three clips.

The clips will be from two baseball games. One set of clips features the 2009 AAA All-Star Game, played between the International League All-Stars and the Pacific Coast League All-Stars on July 9, 2009. The other is a game between the Buffalo Bisons (the New York Mets AAA affiliate) and the Syracuse Chiefs (the Washington Nationals AAA affiliate). This game is from April 7, 2011, the first day of the AAA season that year.

Your rights: Your decision to participate in this study is entirely voluntary and you may decide at any time to withdraw from the study. Your decision to discontinue will not affect your academic status or access to services from the University of Alberta. If you choose to participate, you may skip any items you do not wish to answer. Responses made by individual participants on the questionnaires will remain confidential, and your name will not appear on the questionnaire or be associated with your responses in any way. Questionnaires will be identified only by a researcher-assigned code number, for the purpose of associating survey forms with the particular story that the participant experienced. Only

researchers associated with the project will have access to the questionnaires. The results of this study may be presented at scholarly conferences, published in professional journals, or presented in class lectures. All data presented will be anonymous. The data will be securely stored by (Greg Lee) for a minimum of five years.

Benefits and risks: There are no foreseeable risks to this study, but if any risks should arise, the researcher will inform the participants immediately. If you should experience any adverse effects, please contact Greg Lee and/or Dr. Vadim Bulitko immediately.

Contact information If you have any questions or comments on the study, or if you wish a clarification of rights as a research participant, you can contact Greg Lee or the Human Research Ethics Committee at the number and address below.

Greg Lee
Ph.D. Candidate
27 Finch Court
Kentville, NS
(902) 365-5884

Vadim Bulitko, Ph.D.
Associate Professor
Department of Computing Science
University of Alberta Edmonton, AB
T6G 2E8
(780) 492-3854

University of Alberta Research Ethics Office
(780) 492-2614
reoffice@ualberta.ca

Stephen Maitzen, Ph.D.

Chair of the Acadia Research Ethics Board
Department of Philosophy, Acadia University
Wolfville, NS
B4P 2R6
(902) 585-1407

Signatures. Please sign below to indicate that you have read and understood the nature and purpose of the study. Your signature acknowledges the receipt of a copy of the consent form as well as indicates your willingness to participate in this study.

D.3 Debriefing for User Study I

Thank you for participating in this study! Your time and effort have been valuable to us. Baseball broadcasting is a major business, and we would like to show that part of it can either be automated completely or aided by an automated system. Our research investigates whether computers can learn to do colour commentary for baseball. Specifically, we investigate having computers automatically select stories from baseballs immediate and distant past based on what is actually happening (or has just happened) in the current game. Since storytelling is considered a very human task, this is an interesting and challenging problem.

To examine this, we created an automated storytelling system in which the story is chosen based on features of the current game state (such as balls, strikes and the score). Our independent variable is the presence or absence of the storytelling system within the broadcast. There were three types of broadcast shown in the user study:

- a broadcast with no commentators
- play-by-play commentary with human colour commentary
- play-by-play commentary with computer-chosen colour commentary.

Our hypotheses were that viewers who watched the broadcasts with computer-chosen colour commentary would rate the game's entertainment value as highly as viewers who experienced the broadcasts with human colour commentary. The broadcasts without commentators were included to investigate whether commentary is necessary at all. It was

necessary to withhold the information that the broadcasts could contain computer commentary to eliminate biases based on having this information. The results of this research could help both the video game industry by supplying automatic colour commentary, or actual Major League Baseball broadcasts by supplying an assistant to a human colour commentator.

Thanks very much for participating. Without the help of people like you, we couldn't answer most important scientific questions in psychology. You've been a great help. Do you have any questions that I can answer right now? If you have any questions, later on, about the study, please contact Greg Lee via either phone (902-365-5884) or e-mail (greglee@cs.ualberta.ca) or if you have general questions, contact the University of Alberta Research Ethics Board at reoffice@ualberta.ca or (780) 492-2614 , Sharon Randon (Research Participation Coordinator) at rescred@ualberta.ca or 780-492-5689. or Dr. Stephen Maitzen (Chair of Acadia Research Ethics Board) via phone (902-585-1407) or email (stephen.maitzen@Acadiau.ca). Please do not tell other people about what we had you do here to avoid biasing potential participants.

D.4 Debriefing for User Studies II-VI

Thank you for participating in this study! Your time and effort have been valuable to us. Baseball broadcasting is a major business, and we would like to show that part of it can either be automated completely or aided by an automated system. Our research investigates whether computers can learn to do colour commentary for baseball. Specifically, we investigate having computers automatically select stories from baseballs immediate and distant past based on what is actually happening (or has just happened) in the current game. Since storytelling is considered a very human task, this is an interesting and challenging problem.

To examine this, we created an automated storytelling system in which the story is chosen based on features of the current game state (such as balls, strikes and the score). Our independent variable is the presence or absence of the storytelling system within the broadcast. There were three types of broadcast shown in the user study:

- a broadcast with no commentators
- play-by-play commentary with human colour commentary

- play-by-play commentary with computer-chosen colour commentary.

Our hypotheses were that viewers who watched the broadcasts with computer-chosen colour commentary would rate the game's entertainment value as highly as viewers who experienced the broadcasts with human colour commentary. The broadcasts without commentators were included to investigate whether commentary is necessary at all. It was necessary to withhold the information that the broadcasts could contain computer commentary to eliminate biases based on having this information. The results of this research could help both the video game industry by supplying automatic colour commentary, or actual Major League Baseball broadcasts by supplying an assistant to a human colour commentator.

Thanks very much for participating. Without the help of people like you, we couldn't answer most important scientific questions in psychology. You've been a great help. Do you have any questions that I can answer right now? If you have any questions, later on, about the study, please contact Greg Lee via either phone (902-365-5884) or e-mail (greglee@cs.ualberta.ca) or if you have general questions, contact the University of Alberta Research Ethics Board at reoffice@ualberta.ca or (780) 492-2614 , Sharon Randon (Research Participation Coordinator) at rescred@ualberta.ca or 780-492-5689. or Dr. Stephen Maitzen (Chair of Acadia Research Ethics Board) via phone (902-585-1407) or email (stephen.maitzen@Acadiau.ca). Please do not tell other people about what we had you do here to avoid biasing potential participants.

D.5 Questionnaire for User Study I

Please circle one value for each question:

Video Clips 1-6

I found this viewing experience enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I learned something watching this video clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I found this viewing experience interesting
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I found the video clip easy to follow
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I enjoyed the commentary in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in watching baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in watching sports
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in participating in sports
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

D.6 Questionnaire for User Studies II and III

Please circle one value for each question

Video Clips 1-6

I found this viewing experience enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I learned something watching this video clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I found the video clip easy to follow

(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I enjoyed the commentary in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in watching baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a good balance between the commentators
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the commentators displayed bias towards one team
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

The commentators were experts concerning baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a story in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was fitting to the game at hand
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was annoying
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

D.7 Questionnaire for User Study IV

Please circle one value for each question

Video Clips 1-3

I found this viewing experience enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I learned something watching this video clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I found the video clip easy to follow
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I enjoyed the commentary in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in watching baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a good balance between the commentators
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the commentators displayed bias towards one team
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

The commentators were experts concerning baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a story in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was enjoyable

(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was fitting to the game at hand
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was annoying
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

1) The three clips you have just seen used different forms of commentary. You will now see one more clip from this game. Please choose which form of commentary you'd like to see used in the final clip:

a) Type of commentary from clip 1

b) Type of commentary from clip 2

c) Type of commentary from clip 3

Video Clip 4

I found this viewing experience enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I learned something watching this video clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I found the video clip easy to follow
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I enjoyed the commentary in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in watching baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a good balance between the commentators
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the commentators displayed bias towards one team
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

The commentators were experts concerning baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a story in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was fitting to the game at hand
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was annoying
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I am glad I chose this version of the fourth clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

In hindsight, I would have preferred to choose one of the other versions
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

The type of commentary in this clip was what I expected based on my choice
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

D.8 Questionnaire for User Studies V and VI

Please circle one value for each question
Video Clips 1-6

I found this viewing experience enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I learned something watching this video clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I found the video clip easy to follow
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I enjoyed the commentary in this clip
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

Viewing this clip made me more interested in watching baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

There was a good balance between the commentators
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the commentators displayed bias towards one team
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

The commentators were experts concerning baseball
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

The story in this clip was about:
a) there was no story
b) tomatoes
c) a disputed strike call
d) players trying to throw games

I thought the story in this clip was enjoyable
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was fitting to the game at hand
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

I thought the story in this clip was annoying
(Strongly Disagree) 1 2 3 4 5 6 7 (Strongly Agree)

D.9 Post-Study Questionnaire for User Studies I,II and III

1) Are you a baseball fan?
Yes No

2) Were you familiar with the teams involved in the clips?
Yes No

3) In an average year, how much baseball do you watch? (Please circle one)

None

Less than a full game

1-5 games

5-20 games

20+ games

4) In an average year, how many hours of sports do you watch? (Please circle one)

0 hours

1-10 hours

10-100 hours

100+ hours

5) Please circle your age group:

18-25

26-30

31-35

36 and up

6) Please circle your sex

Male

Female

D.10 Post Study Questionnaire for User Studies IV, V, and VI

1) Are you a baseball fan?

Yes No

2) Were you familiar with the teams involved in the clips?

Yes No

3) In an average year, how much baseball do you watch? (Please circle one)

None

Less than a full game

1-5 games

5-20 games

20+ games

4) In an average year, how many hours of sports do you watch? (Please circle one)

0 hours

1-10 hours

10-100 hours

100+ hours

5) Please circle your age group:

18-25

26-30

31-35

36 and up

6) How many books did you read for pleasure in the past year? (Please circle one)

None

1-3

4-10

10 or more

7) Do you think having stories added to a broadcast makes a game more entertaining?

Yes

No

8) In your lifetime, how many organized baseball/softball games have you participated in? (Please circle one)

None

1-5

6 or more

9) In general, do you think baseball is boring?

Yes

No

6) Please circle your sex

Male

Female

Bibliography

[Allen, 2012] Robbie Allen. Statsheet, 2012. http://www.statsheet.com.

[Altman, 1986] Rick Altman. Television/Sound. *Studies in entertainment: Critical approaches to mass culture*, pages 39–54, 1986.

[Amazon, 2012] Amazon. Mechanical turk, 2012. https://www.mturk.com/mturk/welcome.

[Amini *et al.*, 2008] Massih-Reza Amini, Vinh Truong, and Cyril Goutte. A boosting algorithm for learning bipartite ranking functions with partially labeled data. In *Proceedings of the 31st International Special Interest Group on Information Retrieval (SIGIR) Conference*, pages 99 – 106, 2008.

[Andre *et al.*, 2000a] Elisabeth Andre, Kim Binsted, Kumiko Tanaka-Ishii, Sean Luke, Gerd Herzog, and Thomas Rist. Three robocup simulation league commentator systems. *AI Magazine*, 76:57–66, 2000.

[Andre *et al.*, 2000b] Elisabeth Andre, Thomas Rist, Susanne van Mulken, Martin Klesen, and Stephan Baldes. The automated design of believable dialogues for animated presentation teams. In J. Cassell, S. Prevost, J. Sullivan, and E. Churchill, editors, *Embodied Conversational Agents*, pages 220 – 255. MIT Press, 2000.

[Bailey, 2009] Paul Bailey. Searching for storiness: Story-generation from a readers perspective. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence*, pages 67 – 72, 2009.

[Bryant *et al.*, 1977] Jennings Bryant, Paul Comisky, and Dolf Zillman. Drama in Sports Commentary. *Journal of Communication*, pages 140–149, 1977.

[Bryant *et al.*, 1982] Jennings Bryant, Dan Brown, Paul Comisky, and Dolf Zillman. Sports and Spectators: Commentary and Appreciation. *Journal of Communication*, pages 109–119, 1982.

[Burges *et al.*, 2005] Chris Burges, Tal Shaked, Erin Renshaw, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML*, pages 89–96, 2005.

[Burges *et al.*, 2006] Christopher Burges, Krysta Svore, Paul Bennett, Andrzej Pastusiak, and Qiang Wu. Learning to rank with non-smooth cost functions. In *Proceedings of the Twentieth International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 193 – 200, 2006.

[Burges *et al.*, 2011] Christopher Burges, Krysta Svore, Paul Bennett, Andrzej Pastusiak, and Qiang Wu. Learning to rank using an ensemble of lambda-gradient models. In *Proceedings of the Yahoo! Learning to Rank Challenge*, pages 25 – 35, 2011.

[Burges, 2010] Christopher Burges. From ranknet to lambdarank to lambdamart: An overview. Technical report, Microsoft Research, 2010.

[Chappelle *et al.*, 2009] Olivier Chappelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 621–630, 2009.

[Dix *et al.*, 1993] Alan Dix, Janet Finlay, Gregory Abowd, and Russell Beale. *Human Computer Interaction*. Prentice Hall, 1993.

[Duncan and Hasbrook, 1988] Margaret Duncan and Cynthia Hasbrook. Denial of power in televised women's sports. *Sociology of Sport Journal*, 5:1–21, 1988.

[Esuli *et al.*, 2006] Andrea Esuli, Tiziano Fagni, and Fabrizio Sebastiani. Mp-boost: A multiple-pivot boosting algorithm and its application to text categorization. In *Proceedings of the 13th International Symposium on String Processing and Information Retrieval (SPIRE)*, pages 1 – 12, 2006.

[Frankel, 2012] Stuart Frankel. Narrative science, 2012. www.narrativescience.com.

[Freund and Schapire, 1995] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, 1995.

[Freytag and MacEwan, 1908] G. Freytag and E.J. MacEwan. *Freytag's Technique of the Drama: An Exposition of Dramatic Composition and Art*. Scott, Foresman and Co., 1908.

[Friedman, 1999] J.H. Friedman. Greedy function approximation: A gradient boosting machine. Technical report, Stanford, 1999.

[Hammond *et al.*, 1990] Kenric Hammond, Robert Prather, Visvanath Data, and Carol King. A provider-interactive medical record system can favorably influence costs and quality of medical care. *Computers in Biology and Medicine*, 20(4):267 – 279, 1990.

[Herbrich *et al.*, 2000] Raif Herbrich, Thore Graepel, and Klaus Obermayer. *Advances in Large Margin Classifiers*, chapter Large margin rank boundaries for ordinal regression. MIT Press, Cambridge,MA, 2000.

[Kennedy and Hills, 2009] Eileen Kennedy and Laura Hills. *Sports, Media and Society*. Berg Publishers, 2009.

[Kitano *et al.*, 1997] H Kitano, M Asada, Y Kuniyoshi, I Noda, E Osawa, and H Matubara. Robocup, a challenge problem for AI. *AI Magazine*, pages 73–85, 1997.

[Knoppel *et al.*, 2008] François L. A. Knoppel, Almer S. Tigelaar, Danny Oude Bos, Thijs Alofs, and Zsofia Ruttkay. Trackside DEIRA: a dynamic engaging intelligent reporter agent. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems.*, AAMAS '08, pages 1681–1682, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[Kory, 2012] Matthew Kory. SCOUT, 2012. http://www.baseballprospectus.com/article.php?articleid=16835.

[Lee and Prime, 2008] Bill Lee and Jim Prime. *Baseball Eccentrics*. Triumph Books, 2008.

[Li *et al.*, 2007] Ping Li, Christopher Burges, and Qiang Wu. McRank: Learning to rank using multiple classication and gradient boosting. In *Proceedings of the Twenty First Conference on Neural Information Processing Systems (NIPS)*, pages 845–852, 2007.

[Liu *et al.*, 2008] Tie-yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank - theory and algorithm. In *International Conference on Machine Learning (ICML)*, pages 1192–1199, 2008.

[Live XML game summaries, 2011] Live XML game summaries. Major League Baseball, 2011.

[Manning *et al.*, 2008] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge Press, 2008.

[Michener, 1976] James Michener. *Sports in America*. Fawcett Books, 1976.

[Mitchell, 1997] T. Mitchell. *Machine Learning*. WCB/McGraw-Hill, 1997.

[Moon *et al.*, 2010] Taesup Moon, Alex Smola, Yi Chang, and Zhaohui Zheng. Intervalrank isotonic regression with listwise and pairwise constraints. In *Proceedings of the First International Conference on Web Search and Data Mining (WSDM)*, pages 151 – 159, 2010.

[Neyer, 2008] Rob Neyer. *Rob Neyer's Big Book of Baseball Legends: The Truth, The Lies and Everything Else*. Fireside, 2008.

[Ontanon and Zhu, 2011] Santiago Ontanon and J. Zhu. The sam algorithm for analogy-based story generation. In *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2011)*, pages 67 – 72, 2011.

[Rhodes *et al.*, 2010] Martin Rhodes, Simon Coupland, and Tracy Cruickshank. Enhancing real-time sports commentary generation with dramatic narrative devices. In *International Conference on Interactive Storytelling (ICIDS)*, pages 111–116, 2010.

[Riedl *et al.*, 2008] Mark Riedl, Andrew Stern, Don Dini, and Jason Alderman. Dynamic experience management in virtual worlds for entertainment, education, and training. *International Transactions on Systems Science and Applications, Special Issue on Agent Based Systems for Human Learning and Entertainment*, 3(1):23–42, 2008.

[Riedl, 2004] Mark Owen Riedl. *Narrative generation: balancing plot and character*. PhD thesis, North Carolina State University, 2004. AAI3154351.

[Roberts and Isbell, 2008] David L. Roberts and Charles L. Isbell. A Survey and Qualitative Analysis of Recent Advances in Drama Management. *International Transactions on Systems Science and Applications (ITSSA)*, 4(2):61–75, 2008.

[Ross, 2012] Stephen Ross. FanVision, 2012. http://www.fanvision.com/.

[Ryan, 1993] Marie-Laure Ryan. Narrative in real time: Chronicle, mimesis and plot in the baseball broadcast. *Narrative*, 1(2):138–155, 1993.

[SCE San Diego Studio, 2009] SCE San Diego Studio. MLB 09: The Show, 2009. http://us.playstation.com/games-and-media/games/mlb-09-the-show-ps3.html.

[Smith, 1995] Curt Smith. *Storytellers: From Mel Allen to Bob Costas : Sixty Years of Baseball Tales from the Broadcast Booth*. MacMillan Publishing, 1995.

[Taylor *et al.*, 2008] Michael Taylor, John Guiver, Stephen Robertson, and Tom Minka. Softrank: optimising non-smooth rankmetrics. In *Proceedings of the First International Conference on Web Search and Data Mining (WSDM)*, pages 77 – 86, 2008.

[Thue *et al.*, 2007] David Thue, Vadim Bulitko, Marcia Spetch, and Eric Wasylishen. Interactive storytelling: A player modelling approach. pages 43–48, Stanford, California, 2007. AAAI Press.

[Tsai *et al.*, 2007] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: a ranking method with fidelity loss. In *Proceedings of the 30th International Special Interest Group on Information Retrieval (SIGIR) Conference*, pages 383–390, 2007.

[Volkovs and Zemel, 1999] Maksims N. Volkovs and Richard S. Zemel. Boltzrank: Learning to maximize expected ranking gain. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 1089 – 1096, 1999.

[Williams and Prime, 1996] Ted Williams and Jim Prime. *Ted William's Hit List*. Masters Print, 1996.

[Xu and Li, 2007] Jun Xu and Hang Li. AdaRank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 391–398, New York, NY, USA, 2007. ACM.

[Yoav Freund and Singer, 2003] Robert E. Schapire Yoav Freund, Raj Iyer and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research (JMLR)*, 4:933–969, 2003.

[Young, 2007] R. Michael Young. Story and discourse: A bipartite model of narrative generation in virtual worlds. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, 8(2):177–208, 2007.