StatsMonkey: A Data-Driven Sports Narrative Writer

Nicholas D. Allen, John R. Templon, Patrick Summerhays McNally,
Larry Birnbaum, Kristian Hammond
2133 Sheridan Rd.
Evanston, IL 60208
nallen@narrativescience.com, jtemplon@gmail.com, patrickmcnally2013@u.northwestern.edu,
birnbaum@cs.northwestern.edu, hammond@cs.northwestern.edu *

Abstract

There are certain types of stories that are often told in very structured ways; sports stories or financial reports are two examples. Readers care about these narratives because they are passionately interested in the topic and want to read about the specific details of the event. In other words, they care about the data and want to read a story that presents that data to them. However, in order to be compelling these narratives cannot merely repeat the data, rather they must tell a story from the data. In this paper, we will present a model for data-driven story-telling and discuss StatsMonkey, a system that automatically writes baseball stories from raw baseball game numerical data available online. We will show that a machine can generate interesting, readable stories and that it can make editorial decisions about what aspects of a situation to highlight. Further we will show that a machine can determine in what manner those aspects should be shared.

The StatsMonkey System

Systems like Tale-Spin (Meehan 1977) have been developed to tell stories, but they required the program to have deep domain knowledge and recounted the events in a linear fashion. Conversely, StatsMonkey is a system for writing newspaper-style accounts of sporting events using internet-gathered statistical data from the games. The dataset is primarily play-by-play and box scores. In order to write these stories, StatsMonkey must determine what the narratives apply to a given game, which of those narratives are most interesting or important, and which moments in the game most exemplify each narrative.

While the system currently produces game stories about sporting events, the techniques and technology could eventually be extended to write stories for any data rich domain including stories dealing with finance, crime or the U.S. Census.

The current version of StatsMonkey was developed in part by a working sports-journalist, and the narratives have been built to mimic human sports writing.

In general these narratives fit into three types: extreme or rare events, the crossing of important thresholds and comparisons with historically-based expectations. Because these narratives provide a specific looking glass into the game, but are not comprehensive accounts of the game, we borrow a term from journalism and call them “angles”.

Narrative Angles

StatsMonkey is not designed to list the events of the game, instead it writes the story of the game by emulating what a sports reporter would look for in a game, the “big picture”.

This is achieved through angles, which are arguments for a narrative interpretation of events based on selected data, and are influenced by Roger Schank’s thematic structures (1982). They provide context that is broader than a single element, look for trends in data, explain the outcomes of games, and demonstrate why the reader should care about the story. StatsMonkey attempts to match game events to a wide range of angles.

Each of the above narrative types can aid StatsMonkey in finding angles, but a more powerful analysis of the game comes from combining the factors. For instance, determining a comeback victory in baseball requires an extreme condition (a low chance of victory) followed by the crossing of a threshold (re-taking the lead later in the game). Combining these two narrative concepts gives StatsMonkey a much wider breadth of angles. Currently the system looks for approximately 35 unique angles in the game data.

StatsMonkey determines the importance of events, players and statistics in the game by computing a number of critical derived features above and beyond the raw data. The system takes advantage of three important advanced baseball statistics: Win Probability, Leverage Index and Game Score.

Each of these statistics allows the system to identify key moments and players in a game. Win probability is the historical likelihood of a team winning the game when faced with a certain game state. It allows StatsMonkey to find the plays that impacted the outcome of the game the most. Leverage index is the historical ability for win probability to shift on a given play and it picks out the plays that had the potential for the greatest impact on the game; it also enables StatsMonkey to search for missed opportunities and crucial

[*] Thank you to the National Science Foundation for funding the work on this project through Grant #0535231. Also, thank you to Thu Cung and Tian Huang who worked on the original version of the system, and to Jeremy Gilbert, Rich Gordon, and Owen Youngman for their help and advice.

moments in the game.

Additionally, average leverage index across an inning or game can reflect the character of the event. Finally, game score is a composite metric of a number of game statistics, which are combined to picks out the best hitters and pitchers in the game. StatsMonkey uses an adapted version of game score, which allows it to compare the entire range of baseball players in a game.

In the partial example that follows, which was written entirely by StatsMonkey and appears on BigTenNetwork.com (2010), Michael Earley's performance is selected as the most important narrative, based on a game score that exceeds 75. Digging deeper, the program determines that the most striking aspect of his performance is his two home runs.

> EVANSTON, Ill. – Friday was a great day for Indiana's Michael Earley. He hoisted the Indiana Hoosiers to an 11-9 victory over the Northwestern Wildcats (20-26). Earley blasted two home runs for Indiana (22-21). The senior right field went 3-for-5 in the game with four RBIs and two runs scored. Earley homered in the first and second innings.

Combining pieces of derivative data can provide a very deep analysis of a baseball game. For instance, by combining runners in scoring position, total hits and the inning in which the winning team scored the majority of their runs, the program can write an angle about a winning team that took advantage of fewer hits by bunching them together in one big inning. The combination of the result, leverage index and run differential immediately after the play allows StatsMonkey to write about a clutch pitching performance in an ultimately doomed team effort, or a batter who failed to get a big hit.

Angle Prioritization

It is often the case that StatsMonkey will identify multiple angles for a game. In these situations, the program uses a system designed to mimic a sportswriters intuition about the importance of an event. The elements are ordered on this scale, which is called the priority.

The prioritization of angles is based on a mapping to a 1-to-10 scale of importance. Some angles, such as a pitcher throwing a perfect game, have a single value that they map to, in this case 10. For instance, there have been just 19 perfect games in the history of Major League Baseball. Others angles are scaled based on their key statistical measures. The priority of an angle for a pitcher who fails to accomplish a particular feat of excellence, but still performs well will be determined by a mapping of his game score to an appropriate value on the priority scale.

The Writer

Once the narrative arcs of the story and their associated priorities have been determined StatsMonkey has an outline of the general themes of the game. This outline is then used to generate natural language using dynamic scripts with slots for the relevant data points. Like a sportswriters recap would, StatsMonkey begins with the most important angle (the one with the highest priority) first. This first angle generally receives a more detailed treatment; hence the importance of priority, with the subsequent angles merely adding other pieces of important information. Because StatsMonkey has an outline of the relevant players and moments in the game it could also be used to drive video highlights or writing in a language other than English. The outline can also be used to generate queries to search for other supporting information on the web, including quotes and pictures from the game.

References

Earley powers Indiana past Northwestern. 2010. http://www.bigtennetwork.com/dpp/sports/baseball/Earley-powers-Indiana-%past-Northwestern.

Meehan, J. R. 1977. Tale-spin: An interactive program that writes stories. *In Proceedings of the Fifth International Joint Conference on Artificial Intelligence*.

Schank, R. C. 1982. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Addison-Wesley.