Automated Sports Journalism

Eero Laine
Department of Computer Science
University of Helsinki

Automated journalism is a relatively new technological innovation where actual news content is created using natural language generation (NLG) [1]. Sports journalism has been considered an ideal domain for automation because of the wealth of available statistics and the use of formulaic templates and style in sports reporting [2]. Baseball has been described as the ideal sport for automated journalism because of the available data, statistical focus and advanced predictive models [1]. The leading natural language generation technology companies in the United States, Automated Insights and Narrative Science, were born from attempts to generate game recaps automatically [1]. Associated Press (AP) attracted wide media coverage when it announced its collaboration with Automated Insights to provide automated game recaps for Little League Baseball [3], [4], [5], [6].

New Yorker magazine [3] describes a formula for game recaps, where the first paragraphs provide the final score and the turning point of the game. Unimaginative postgame citations and game statistics are repeated throughout the recap. If vital statistics presented in a short, readable form is all that is required, an automated recap may well suffice, New Yorker suggests. Editor Jason Fry argues game recaps have become less important and automated recaps can be "good enough" despite their obvious deficiencies [7].

Automated journalism requires data in a machine-readable format, like a spreadsheet [1]. The Guardian newspaper [8] refers to the automated content as data-driven, also noting how the available data limits the content.. AP executive Barry Bedlan emphasizes the accuracy of data received from a data provider: "...if a sports organization cannot give us 100% stamp of approval on accuracy for hours or even days that does not work, it has lost its news value for a newspaper, broadcaster or website" [4]. According to former AP executive Lou Ferrara, data for AP's Minor League recaps was entered by coaches without strict verification, resulting in errors [1]. When Narrative Science provided their own automated recaps of Little League games, the company relied on data provided by volunteers [9].

The deficiencies of automated game recaps have been extensively covered. Sports journalists' knowledge is difficult to apply in an automated format [1]. Narrative Science's chief scientist Kris Hammond argues the challenge of automated sports journalism is in setting the right context for the statistics and finding the most important facts from them [10]. This importance of finding the meaning of statistics and putting them in the right context is also emphasized by other writers [7], [8]. Sports Journalism Professor Malcolm Moran criticizes automated recaps for failing to provide the sports fans the "insider" details: "The story of the day may have much more to do with off-the-field considerations than whether somebody went 0-for-5" [7]. The interviews made by sports reporters are missing from the automated recaps, which some journalists consider a major weakness [3], [6] .

The journalistic style of automated recaps has been unfavorably compared with that of real writers, whose language is more colorful [3], [6]. Automated Insights has attempted to imitate the language used in human-written sports journalism, and the results have been referred as clichéd and formulaic [7], [11]. In return, Automated Insights CEO Robbie Allen has openly derided the quality of sports journalism [3], [8], [12].

Cost and speed are two factors where automated recaps are clearly better than human sportswriters [3]. AP did not have resources to cover Minor League Baseball before the news agency's collaboration with Automated Insights [4]. Dörr [13] argues automated journalism can provide profitable coverage for special interest topics. However, we could find no sources for the actual readership of automated game recaps currently readable.

Automated Insights' NLG product Wordsmith is not available in public, but it appears to be template-based [14], [15]. Wordsmith's user-directed documentation [15] presents user-created templates, which can be programmed to show different data while keeping the core structure. Wordsmith features branches, essentially conditionals which change the text's structure based on the data. The user can also pre-set synonyms to give the text variation without changing the formula. If this is Wordsmith's core functionality, we suggest the product's success may not lie in any advanced algorithm use, but an accessible graphic user interface. However, it is unclear to us how much Automated Insights tailors its solution to a high-profile customer such as AP.

Narrative Science's first prototype, baseball text generator StatsMonkey [1], appears to be more ambitious in its NLG methodology. StatsMonkey is based on narrative angles, which are "arguments for a narrative interpretation of events based on selected data", which look for specific trends from the data and provide the necessary context for it [16]. Based on three core baseball statistics and historical data, StatsMonkey attempts to provide analysis, not just a formulaic recap. We suggest these sorts of analytical tools may prove to be most useful in the future, while template-based sports journalism is limited by the strong role humans play in its execution.

References

[1] Graefe, A. (2016). Guide to automated journalism.
[2] Van Dalen, A. (2012). The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. Journalism Practice, 6(5-6), 648-658.
[3] http://www.newyorker.com/news/sporting-scene/the-sportswriting-machine
Retrieved 1 January 2017
[4]
https://techcrunch.com/2016/07/03/ap-sports-is-using-robot-reporters-to-cover-minor-league-baseball/?ncid=rss Retrieved 2 February 2017
[5]
http://www.theverge.com/2016/7/4/12092768/ap-robot-journalists-automated-insights-minor-league-baseball Retrieved 2 February 2017
[6] http://www.bbc.com/news/technology-34204052 Retrieved 2 February 2017
[7]
http://www.poynter.org/2010/statsheet-technology-generates-game-stories-with-surprising-insights-unsurprising-cliches/111565/ Retrieved 1 February 2017
[8]
https://www.theguardian.com/small-business-network/2016/jul/22/written-out-of-story-robots-capable-making-the-news Retrieved 2 February 2017
[9] Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. Journalism Practice, 8(5), 519-531.
[10] Wright, A. (2015). Algorithmic authors. Communications of the ACM, 58(11), 12-14.
ISO 690
[11] Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. Digital Journalism, 3(3), 416-431.
[12]
http://www.poynter.org/2010/statsheet-network-automates-hundreds-of-sports-stories-from-databases/108463/ Retrieved 2 February 2017
[13] Dörr, K. N. (2016). Mapping the field of Algorithmic Journalism. Digital Journalism, 4(6), 700-722.
[14] http://www.wired.co.uk/article/wordsmith-robot-journalist-download Retrieved 2 February 2017
[15] https://wordsmithhelp.readme.io/ Retrieved 3 February 2017
[16] Allen, N. D., Templon, J. R., McNally, P. S., Birnbaum, L., & Hammond, K. J. (2010, November). StatsMonkey: A Data-Driven Sports Narrative Writer. In AAAI Fall Symposium: Computational Models of Narrative.