Journalist versus news consumer: The perceived credibility of machine written news

Hille van der Kaa
Fontys University of Applied Sciences
Tilburg University
The Netherlands
h.vanderkaa@fontys.nl

Emiel Krahmer
Tilburg center for Cognition and Communication
Tilburg University
The Netherlands
e.j.krahmer@uvt.nl

ABSTRACT

This research aims to contribute to the unexplored field of audience studies with a focus on the credibility of automated journalism. In this paper, we take a systematic look into the perceived credibility of robot-written news articles, searching specifically for differences and similarities between journalists and news consumers. In total, 232 native Dutch speakers (the language of the experiment) took part in this research, and among them were 64 journalists. The participants were asked to evaluate the perceived levels of the expertise and trustworthiness of four news articles based on algorithms outlined in the data-to-speech system (D2S) and created by Theune et al. (2001). We used a 2 (author: computer or journalist) x 2 (story topic: sport or finance) between-subject design to determine the perceived credibility of the news writer (source) and the contents of the news story (message).

Within the group of news consumers, no main effect was found. News consumers perceived the levels of the trustworthiness and expertise of the computer writer and journalist equally. Within the group of journalists, we found a significant effect on the perceived trustworthiness of the news source. In our experiment, journalists perceived the trustworthiness of a journalist to be much higher than that of the computer. Further, journalists perceived the expertise of the computer to be higher than the news consumers perceived it to be. Finally, the story topic has an influence on the item's perceived trustworthiness. Overall, respondents perceived the trustworthiness of a sports article to be lower than that of a finance article.

It will be interesting to investigate this topic further, as it is possible that these differences between journalists and consumers will increase along with the rise in automated storytelling.

Keywords

Credibility – Automated Journalism – Natural Language Generation – Computer Written News – Robot Journalism

INTRODUCTION

The automation of journalism has entered a new phase with the rise of computer-written news articles. Although template-based text-generation algorithms have been under development since the 1990s (e.g., Buseman & Horacek, 1998; McRoy, Channarukul, & Ali, 2003; Narayan, Isbell, & Roberts, 2011; Theune, Klabbers, Odijk, De Pijper, & Krahmer, 2001), news companies have only recently started to work with algorithms operating along similar lines (e.g., Nichols, Birnbaum, & Hammond, 2014). These increasing technological developments have led to a new type of journalism: robot journalism (Van Dalen, 2012). Graig Silverman, an award-winning journalist and Adjunct Faculty at the Poynter Institute[1], believes these new techniques will improve the levels of accuracy and quality of journalistic reports. Van Dalen (2012) adds that journalists see the advantages of an algorithm as an unbiased reporter. However, does the general audience agree?

Previous research on the evaluation of natural language generation (NLG) systems often focused on the quality of the generated text. Research on the levels of credibility of human-written news articles is abundant in the field of Media Studies. Further, the way journalists re-examine their core skills in the light of new technological developments is well discussed by Singer (2005) and Van Dalen (2012), amongst others. However, audience studies in the era of the credibility of automated journalism are still extremely thin on the ground. As far as we are aware, the only previous study that addressed this is one recently published by Clerwall (2014). Building on this "initial small-scale study" (in the words of the author), we look more systematically into the perceived credibility of robot-written news articles, searching specifically for differences and similarities between journalists and news consumers in how they rate computer-generated news.

Computer writers and journalists: what they have in common

News writing relies on a basic formula, and there are key elements to every news story, as follows. The first step is to look for a news occasion. Strong stories are backed by proper *research*, not the exhaustive kind a scientist would conduct, but enough for the journalist to understand and convince their reader of their angle or story point (Synge, 2010). The next step is to *select* the main elements of the story; determining what details to provide to a reader and when are key. Before writing the news story, these details must be *structured* and put in the right order. The last step is to *write* the story in actuality, starting with the lead (Harrower, 2010).

The key elements of the work of a journalist show strong similarities with the tasks of a robot writer. 'Robot writing' or 'algorithmic writing' comes from the field of NLG, which is the process of automatically creating natural language text based on non-linguistic input (e.g., Reiter & Dale, 1997, 2000). It combines knowledge about language and the application of the domain to produce documents and reports automatically. Typically, an NLG

[1] http://www.poynter.org/latest-news/regret-the-error/205816/5-ways-robots-can-improve-accuracy-journalism-quality/

system must be able to perform a number of standard tasks (e.g., Mellish et al., 2006). Interestingly, these tasks are much in line with the previously mentioned basic formula of news writing. According to Reiter and Dale (1997), an NLG system should first determine which information is expressed ('*research,*' often called 'content selection' in the context of NLG). Second, it must organize the available information and determine a structure for the text ('*selecting,*' or 'text planning' in NLG terms). Furthermore, it should determine which information is placed in any sentence and should choose the right words to express the information in the right way ('*structuring*' or 'sentence planning'). Finally, it must create expressions to display, as well as grammatical sentences. This set of tasks ultimately leads to a grammatically correct and clear text ('*writing*'/'linguistic realization').

The first NLG systems produced very simple texts with little or no variation; however, over the years, more linguistic insights were included and techniques were developed for generating more varied texts (see e.g., van Deemter et al., 2005, for discussion).

Nowadays, companies including Narrative Science, Automated Insights, Yseop, and CBS Interactive, as well as start-ups like Fantasy Journalist, create computer-written texts, which are far more profound and 'human-like' than the first NLG systems. It is even hard to see the difference between human-written and computer-written articles, as concluded by Clerwall (2014).

Can robot writers replace journalists?

If the difference between the journalistic content produced by a piece of software and the content produced by a journalist is not evident, will robots replace journalists? A large part of this decision is based on economics: is it cheaper to hire and train a person or to create and maintain a software? The economic decision will usually depend largely on the volume of text produced (Reiter & Dale, 1997).

Clerwall (2014) adds that robot journalism will free resources, allowing reporters to focus on assignments they are more qualified for and leaving the descriptive summaries to the software. He quotes Flew et al. (2012) to strengthen this positive outlook: "Ultimately the utility value of computational journalism comes when it frees journalists from the low-level work of discovering and obtaining facts, thereby enabling greater focus on the verification, explanation and communication of news." According to Van Dalen (2012), journalists tend to take this approach to robot journalism as well. They consider robot journalism an opportunity to make journalism more human. When routine tasks can be automated, journalists will have more time for in-depth reporting.

Cost is not the only factor, however. In some cases, it is technically impossible to generate the required texts with current NLG technology. In these cases, manual production is the only option. Van Dalen (2012) states that machines do not have the creativity to avoid clichés and add humor, although others might disagree (e.g., Binsted & Ritchie, 1994; Petrovic & Matthews, 2013). Moreover, they do not have the flexibility, so some claim, to write non-routine stories, and they do not have analytical skills to offer their stories.

On the other hand, in some cases, NLG techniques may be preferred over manual document creation because they increase accuracy and reduce updating time, or because they guarantee conformance to standards. In addition, NLG techniques enable personalized news presentations, wherein different variants of the same news text can be generated with specific audiences in mind (e.g., variants for skilled as well as low-literacy readers, see e.g., Wubben, van den Bosch, & Krahmer, 2012). Finally, Van Dalen (2012) adds that journalists see the advantages of an algorithm as an unbiased reporter.

However, does the audience agree? Does the audience think automatically generated news stories are as trustworthy as human-written articles?

Perceived credibility of human- and machine-written news stories

Previous research on the evaluation of NLG systems (see e.g., Bangalore, Rambow, & Whittaker, 2000; Belz & Gatt, 2008; Mellish & Dale, 1998) has often focused on the quality of the generated text—is it grammatically correct and coherent? Does it express the required information in the appropriate manner? In the case of domain-dependent template-based text generation systems, grammar issues are perhaps of less importance. These systems are designed in such a way that all relevant information is expressed in grammatical texts. However, other issues do arise, such as how reliable and credible the generated text—in our case, a robot-written news story—appears to be.

Credibility is at the heart of every assessment of a news story (Clerwall, 2014). Previous research on human-written news articles can be grouped into three different approaches: source credibility (Hovland & Weiss, 1951), message credibility (Hamilton, 1998), and medium credibility (Kiousis, 2001; Miller & Kurpius, 2010). Hovland and Weiss (1951) set the starting point for credibility studies. According to them, credibility can be seen as a universal characteristic of a general communication source (e.g., a person or an organization), as based on the dimensions of expertness and trustworthiness. Over the years, other scholars added several variables. Hamilton (1998) highlights the importance of the receiver's decision needs and message topic. Message credibility is explored by focusing on the characteristics of messages that could make them more credible. According to Miller and Kurpius (2010) and others, the credibility of each message is directly influenced by the medium in which it appears.

Fogg and Tseng (1999) state that in discussing the credibility of a computer product, one is actually discussing the perception of the credibility. According to them, credibility can be defined as believability. Credible people are believable people. Credible information can be seen as believable information. Trustworthiness and expertise are hereby two key components. Trustworthiness holds the perceived goodness or morality of a source. Expertise is defined by knowledge, experience, and competences. Taken together, these insights suggest that highly credible computer products will be perceived as having high levels of both trustworthiness and expertise (Fogg & Tseng, 1999).

In this paper, we take a more systematic look into the perceived trustworthiness and expertise of robot-written news articles, searching specifically for differences and similarities between journalists and news consumers.

Experiment

Participants

In total, 232 native Dutch speakers (the language of the experiment) took part in this research, and among them were 64 journalists. Of the respondents, 45.7% were male and 54.3% were female. The youngest to take the survey was 19 and the oldest was 65.

Materials and procedure

Four Dutch robot-written news items were created, which were based on algorithms outlined in the data-to-speech system (D2S) and created by Theune et al. (2001). This system is able to generate simple and straightforward news articles that are similar to basic news articles written by press agencies like Reuters. Two of the articles reported on a sports event (results of a football match), and the other two articles considered a finance topic (stock prices) (see appendix for example text). The contents and sentences were the exact same for both topics, and only the source (computer or journalist) was manipulated. The manipulations, 'this article is written by a computer' or 'this article is written by a journalist,' were shown on every page.

Each participant was randomly presented with one of the four texts. Participants were asked to evaluate the perceived expertise and trustworthiness of the news writer (source) and of the contents of the news story (message). Expertise and trustworthiness were evaluated based on 12 items (translated from Dutch: expertise, intelligence, education, trustworthiness, authority, bias, accuracy, completeness, fact-based, text quality, honesty) using a 5-point Likert scale (ranging from 1 = very low to 5 = very high). We used a factor analysis to compute two valid scales related to expertise (including expertise, intelligence, and authority) and trustworthiness (including reliability, honesty, accuracy, and fact-based).

Design and statistical analysis

In our experiment, a 2 (author: computer or journalist) x 2 (story topic: sports or finance) between-subject design was used. To test for statistical significance between the journalists and news consumers regarding perceived credibility, an analysis of variance (ANOVA) was conducted.

Results

First, we looked at the perceived credibility of the news source (computer versus journalist). Within the group of news consumers, no main effect was found. News consumers perceive the trustworthiness and expertise of the computer writer and journalist equally. In general, they were neutral about the levels of expertise of both the computer writer and journalist writer. News consumers valued the perceived trustworthiness of a computer writer, though they were slightly negative about the expertise of a journalist writer.

Within the group of journalists, there was no effect on the perceived expertise of the news source. In general, journalists were slightly positive about the levels of expertise of the computer writer and the journalist writer. We found a significant effect on the perceived trustworthiness of the news source ($F [1,60] = 7.48$, $p = .008$). In our experiment, journalists perceived the trustworthiness of a journalist as greater ($M = 2.48$, $SD = .11$) than the computer ($M = 2.00$, $SD = .18$).

Journalists and news consumers did not differ in their perceived levels of trustworthiness between machine writers and journalist writers. However, they do differ in their perceptions of expertise ($F [1,166] = 1.90$, $p = .048$). Journalists perceived the expertise of a computer as greater ($M = 3.49$, $SD = .13$) than it was perceived by news consumers ($M = 3.19$, $SD = .10$).

After the analysis of the differences between the perceived credibilities of journalists and news consumers, we measured the influence of the story. The story topic influenced the item's perceived trustworthiness ($F [1,160] = 14.49$, $p = .000$). Overall, the respondents perceived the trustworthiness of a sports article as lower ($M = 2.08$, $SD = 0.58$) than that of a finance article ($M = 2.41$, $SD = .066$).

Discussion

In this research, we found no differences in the perceptions of news consumers regarding the credibility of machine-written news articles.

Therefore, this research lends further support to the pilot findings by Clerwall (2014). Although the method of this current research was somewhat different (participants in this case were fully aware of the fact that they were reading a computer-written article), the idea that news consumers have no strong negative or positive feelings toward computer-written news is still strengthened.

They differ from journalists in this respect. Journalists perceive themselves as more trustworthy compared to their 'computer colleagues.' It will be interesting to investigate this topic further, as it is possible that the differences between journalists and consumers will increase along with the rise of automated storytelling.

Journalists differ from news consumers in their evaluations of the levels of expertise. Journalists perceive the expertise of a computer as higher than it is perceived by news consumers. The evaluation of this specific aspect can be influenced by a more general, positive first impression of a machine-written article. Several journalists left remarks, such as, 'This is actually not bad for a computer.'

Finally, the differences in the perceived levels of trustworthiness of sports and finance articles indicate that further research should take into account other variables as well. Our suggestion would be to further investigate the influence of the story topic, the opinioned stories, and the story type (hard versus soft news), amongst others.

Acknowlegements

We thank Carel van Wijk for his help with the statistical analyses.

References

[1] Bangalore, S., Rambow, O., & Whittaker, S. (2000). Evaluation metrics for generation. In *Proceedings of the 1st International Conference on Natural Language Generation* (pp. 1–8). Mitzpe Ramon.

[2] Belz, A., & Albert, G. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Columbus, OH.

[3] Binsted, K., & Ritchie, G. (1994). *An* implemented model of punning riddles. In *Proceedings of the twelfth national conference on Artificial intelligence* (vol. 1), 633–638. Menlo Park, CA, USA.

[4] Busemann, S., & Horacek, H. (1998). A flexible shallow approach to text generation. In *Proceedings of the 9th International Workshop on Natural Language Generation* (pp. 238–247). Canada.

[5] Clerwall, C. (2014). Enter the Robot Journalist. *Journalism Practice, 8*(5).

[6] Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. *Proceedings of CHI'99, Human Factors in Computing Systems* (pp. 80–87).

[7] Hamilton, M. A. (1998). Message variables that mediate and moderate the effect of equivocal language on source credibility. *Journal of Language and Social Psychology, 17*, 109–143.

[8] Harrower, T. (2010). *Inside Reporting. A practical guide to the craft of journalism.* New York: McGraw-Hill.

[9] Hovland, C., & Weiss, W. (1951). The influence of source on communication credibility effectiveness. *Public Opinion Quarterly, 15*(4), 635–650.

[10] Kiousis, S. (2001). Public Trust or mistrust? Perceptions of media credibility in the information age. *Mass Communication & Society, 4*(4), 381–403.

[11] McRoy, S. W., Channarukul, S., & Ali, S. S. (2003). An augmented template-based approach to text realization. *Natural Language Engineering*, *9*(4), 381–420.

[12] Mellish, C., & Dale, R. (1998). Evaluation in the context of natural language generation. *Computer Speech and Language*, *12*, 349–373.

[13] Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A reference architecture for natural language generation systems. *Natural Language Engineering, 12*, 1–34.

[14] Miller, A., & Kurpius, D. (2010). A Citizen-Eye View of Television News Source Credibility. *American Behavioral Scientist*, *54*(2), 137–156.

[15] Narayan, K. S., Isbell, C. L., & Roberts, D. L. (2011). DEXTOR: Reduced Effort Authoring for Template-Based Natural Language Generation. In *Proceedings of the Seventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 170–175).

[16] Nichols, D. N., Birnbaum, L. A., & Hammond, K. J. (2014). *U. S. Patent No. 8,630,844*. Washington, DC: U. S. Patent and Trademark Office.

[17] Petrovic, S., & Matthews, D. (2013). Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria.

[18] Reiter, E., & Dale, R. (1997). Building applied natural language generation. *Natural Language Engineering*, *3*(1), 57–87.

[19] Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge University Press.

[20] Singer, J. (2005). The political j-blogger. 'Normalizing' a new media form to fit old norms and practices. *Journalism, 6*(2), 173–198.

[21] Synge, D. (2010). *The survival guide to journalism.* Maidenhead: McGraw-Hill.

[22] Theune, M., Klabbers, E., De Pijper, J. R., Krahmer, E., & Odijk, J. (2001). From Data to speech: a general approach. *Natural Language Engineering, 7*(1), 47–86.

[23] Van Dalen, A. (2012). The algorithms behind the headlines. *Journalistic practice, 6*(5–6).

[24] Van Deemter, K., Krahmer, E., & Theune, M. (2005). Real vs. template-based natural language generation: A false opposition? *Computational Linguistics*, *31*(1), 15–23.

[25] Wubben, S., van den Bosch, A., & Krahmer, E. (2012). Sentence Simplification by Monolingual Machine Translation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 1015–1024. Jeju, Korea.

Appendix

Example [Dutch].

A sports news article based on algorithms outlined in the data-to-speech system (D2S), created by Theune et al. (2001). Note: only Dutch articles were used in this experiment.

RKC Waalwijk speelt gelijk tegen PEC Zwolle

RKC Waalwijk ging op bezoek bij PEC Zwolle en speelde gelijk. Het duel eindigde in 1–1. Twaalfduizend toeschouwers kwamen naar het IJsseldelta stadion.

De ploeg uit Zwolle nam na 44 minuten de leiding door een treffer van Saymak. Na achtenveertig minuten bracht Joachim van RKC Waalwijk de teams op gelijke hoogte.

De wedstrijd werd gefloten door scheidsrechter Kamphuis. Hij deelde geen rode kaarten uit. Tomas en De Boer van PEC Zwolle en Sno van RKC Waalwijk liepen tegen een gele kaart aan.

Dit bericht is geschreven door een computer

(Translation example 1 – not used in experiment).

RKC Waalwijk versus PEC Zwolle: match ends in a draw

RKC Waalwijk visited PEC Zwolle and drew. The duel ended in one – all. Twelve thousand spectators came to the Ijsseldelta stadium.

The team from Zwolle took the lead after 44 minutes with a goal by Saymak. Four minutes later, Joachim from RKC Waalwijk equaled the score.

The match was officiated by referee Kamphuis. He did not issue any red cards. Tomas and De Boer of PEC Zwolle picked up a yellow card.

This article was written by a computer.