# The Bright Future of News Automation

Carl-Gustav Lindén

Swedish School of Social Science, University of Helsinki and Sdertrn University
carl-gustav.linden@helsinki.fi

## 1 Introduction

Sometimes, bad metaphors can cause trouble. The idea that robot journalists are coming to take the jobs in newsrooms are getting traction [10] [8]. A picture search produces illustrations of robots banging away at keyboards. However, it should not come as a surprise to anyone that there are no robots moving around. Instead, we are talking about new software introduced in the newsroom, a process that already has been going on for decades [13]. That is the background for this paper with the aim to present some experiences of news automation, identify some important obstacles and explain why the future is still bright. News automation can mean many things, but here we refer to Natural Language Generation (NLG) and automated generation of articles, from numbers to text. That is a much smaller field within language technologies compared to Natural Language Processing (NLP), text to numbers. For an overview of NLG, see [17] and specifically for research into news automation [17] [12] [11]. The media industry has just taken the first small steps on a journey towards a more advanced software environment with systems that will help journalists perform tasks they cannot manage  too much data - and work they do not want to do. Focusing the discussion on the metaphor robot journalist is detrimental to this positive development. We should stop talking about robots when it comes to making stories on news produced by software. Most of the work, writing text templates, is pure handicraft where journalists very much are in charge [14]. If there were any mechanical parts involved, the metaphor should not be a humanoid, but the non-intrusive and useful washing machine. Just like sorting clothes according to fabric and colour, journalists clean – select and sort – the data, dump it in the washing machine, choose the right programme and push the start button. The washing machine metaphor does not evoke imagined existential threats to jobs. Another word of caution is needed. Journalists, researchers and service providers should also stop calling NLG news automation systems smart or AI for now. There is no intelligence here because these systems can only say what happened but have no clue when it comes to the why something happened [21]. Automated systems can report a figure, but they cannot yet say what it means; on their own, computer-generated stories contain no context, no analysis of trends, anomalies, and deeper forces at work. Machine learning or neural networks, for instance, are still far away from newsroom reality.

## 2 Historical background

Already the Surrealist movement experimented with automated drawing and writing, automatism. However, news automation as we know it was actually for the first time conceptualised in a debut novel from 1965, The Tin Men by Michael Frayn [5]. The satire is about the lives of workers at the fictitious William Morris Institute for Automation Research, a British think tank trying to develop robots that are supposed to relieve the employees from work. It is filled with bizarre characters and schemes. In the book, Dr. Goldwasser is working on the automation of newspapers and has invented this fictive language technology UHL (Unit Headline Language) for that purpose. He had collected multi-purpose monosyllables used by headline-writers, such as fear, ban, dash, strike, and fed them into a computer. UHL takes standard headline words and creates headlines from them. These newspaper headlines are completely meaningless. Examples such as "Strike Threat Probe" and "Lab Row Looms" produce sentences - Headline Pidgin - that everybody recognises but nobody can explain the meaning of. Dr. Goldwasser invented three different ways of creating headlines with UHL. By adding one unit at random to the formula each day, or cumulatively or entirely at random you get new headlines. A decade later, in the late 1970s, Tale-Spin, a conceptual NLG program that writes stories, was introduced 10. In the mid 1980s, German researchers created a system designed to produce German newspaper stories about labor market developments, with the somewhat unfortunate name Semtex [18]. And the FoG weather-report generator, the first-ever deployed operational NLG system, goes back to the early 1990s [6]. Frayns novel was published more than 50 years ago and to people who have studied the topic and been involved in R&D, in the darkest moments it seems that there is not much development beyound the properties of UHL. It is not that difficult to generate sentences and texts that at first sight seems to work quite well but after adding complexity to the system, that is, more data and correlations creates unsolvable problems. Looking at the state of the art in news automation, the company that comes first into mind is Narrative Science, created as the student project StatsMonkey at Northwestern University in Chicago 2008-2009 [1], patented and commercialised by two professors of computer science, Kris Hammond and Larry Birnbaum. It was recently selected the most innovative company in Chicago. Narrative science started by providing sports texts, but the company quite soon realised that news media is not the best customer, so they are now serving other business areas, mainly in the financial sector. This seems to be quite common in the NLG business and should make the media sector very worried. Ehud Reiter, a NLG founding father and professor in Aberdeen, also involved as chief scientist in the British leading NLG company Arria, made a small study and found out that the most important sector for commercial NLG is in business reporting. This is perhaps because (a) there is a lot of money in finance, and (b) data and use cases are similar enough to allow systems to be replicated. [16] News media is not even mentioned in the analysis. Against this backdrop, it might seem bold to claim that the future of news automation is bright. One critical issue is the willingness of journalists to be involved in

creating NLG systems and their own perceived roles, especially when it comes to professional ethics. However, this discussion can be considered an extension of the already existing online challenge on professionalism [20]. So why should journalists get involved and in what way can journalists add value? These five points are the main reasons: 1) truth and accuracy, 2) independence, integrity, 3) fairness and impartiality, 4) humanity, and 5) accountability. In short, journalists need to make sure that the outcomes of NLG development are ethically, morally and socially tolerable. We can probably think of many other reasons, such as the brutally obvious probability that engineers and their managers will do this anyway if journalists do not want to be involved.

## 3 The ethical dimension

For some practical guidance, let us turn to the Society of Professional Journalists in the United States and their SPJ Code of Ethics. There we find five central rules where ethical dimensions of news automation are most the obvious. Here is a brief explanation on how to translate ethics into preferred action.

- Journalists should take responsibility for the information they provide, regardless of medium This means that journalists should be willing to get involved in how news automation is designed, where the data comes from, how it is relevant from a journalistic point of view and what kind of decisions algorithms make. There might be some obstacles here, such that newsrooms are short of journalists trained in computational thinking and able to work with programmers. Further, these NLG applications are often software systems bought from a service provider and practically black boxes to journalists.
- Identify sources clearly. The public is entitled to as much information as possible to judge the reliability and motivations of sources The media should be accountable and transparent with where the data comes from and how it is processed. Journalists are supposed to explain ethical choices and processes to audiences. Is it possible or necessary to start a dialogue with the public about journalistic practices, coverage and news content when it comes to news automation? Is the public actually that interested?
- Provide access to source material when it is relevant and appropriate Here there might be a problem that data sources or the software is proprietary and owned by an external commercial company. Media companies might also be afraid to open up processes too much since inputs to algorithms could be gamed by, for instance, public relations experts.
- Avoid stereotyping. Journalists should examine the ways their values and experiences may shape their reporting. We need to ask how values are built into these systems and what are the biases in how news is personalised, for instance gender, race or religion. There is an understanding among media companies that certain types of biases should be avoided, but no clear consensus or written down rules.

– Acknowledge mistakes and correct them promptly and prominently. Explain corrections and clarifications carefully and clearly. Journalists should be able to explain errors that come from faulty data or wrong assumptions built into to news automation systems. Explainable machines in algorithmic decision-making has a lot of appeal [19]. The news agency Associated Press has actually solved this problem by built in features that allows the engineers to backtrack decisions made by the system and check where it went wrong. These features are not shared with journalists in the AP newsroom. In The EU General Data Protection Regulation GDPR it is said that owners of automated decision systems should be able to explain how the system works, they cannot have black boxes. What this means in practice is a bit unclear to everyone as we are all eagerly waiting for cases to end up in court.

Transparency and accountability are two key features of responsible professional journalism. Stuart Myles, a manager at the US news agency AP, has proposed [15] a way to discuss transparency that might be useful. The model consists of four layers:

– The minimum type of algorithmic news transparency is disclosure and by that, Myles means making it known that algorithms have been used in creating or making decisions about news items.
– A step up from disclosure would be justification. A justification aims to show that the results of the algorithmic news are reasonable in a particular instance.
– An explanation is a more comprehensive form of transparency than a justification. It indicates why a particular decision, categorization or arrangement of news was selected and not some other.
– By reproduction Myles means providing sufficient information that the results of a news algorithm could be independently replicated. News organizations sometimes provide the underlying data and algorithms, which they used in a particular report, with the goal of making their use of news algorithms transparent. This feature comes quite close to the concept of open source.

## 4 Why work with journalists?

Few computer or data scientists have a history of working with journalists. Here are some general stereotypes of their professional identity: Journalists generally do not think in structures, they are preoccupied with stories and narratives, they are creative with words and hate repetition, which can become a hindrance in mass production of texts, but also believe they are bad at math. Journalists usually want to be in charge when working with other professional groups and they often tend to be difficult people, nay-sayers with a preference for asking critical questions 3. Actually, media work is a form of affective labour, which is passionate: extreme emotions are part of making the work meaningful [4]. At the same time, journalism and news media is risk-averse due to the nature of their

industry that traditionally have been characterised by deadline thinking, getting the news delivered on a tight schedule. For computer scientists, maybe the most problematic feature is the inability of journalists to explain central concepts in their profession, such as what is news [7]. But in NLG development, journalists need to work in teams with computer scientists for several reasons: their inability to handle large diversity in news input or output with data sources of high dimensionality or high degrees of personalisation. They also aim for maximum impact and do not want to serve micro audiences with news, which is possible at small marginal cost with NLG. How can these two groups work together? Yes, definitely, and media business scholar Lucy Kng who has studied Silicon Valley modes of organisations and culture and compared that to legacy media think [9] they have at least three things in common: 1) Journalists and engineers love the clarity of thought, whether it is as code or as language, 2) they have a strong commitment to the craft, and 3) they share high intrinsic motivation, a sense of purpose.

## 5 Conclusions

So, at last, what is the bright future of news automation? Our own research group Immersive Automation has written an industry report for the newspaper publishers association WAN-IFRA (forthcoming 2019). One of the chapters deals with the future and besides our own knowledge, we asked experts on their views. This is the result of the query. The future of news automation will be about two parallel processes: 1) Decomposition, or deconstruction, of the fundamental principles of journalism 2) Breaking down journalistic work into the actual information artefacts and micro processes The most important question to ask here is What can be automated and what are inherently human tasks? Further, access to open data of high quality is a key issue. Without interesting and dynamic data feeding the NLG systems to produce updated versions of texts with new information and new angles depending on the storyfinding process. There will also be more flexible NLG systems, which makes it easier and cheaper to develop versions for chat bots or talking/listening machines. With the implementation of updated mobile networks, namely 5G, we will see new opportunities for creating and distributing immersive and meaningful content text, video, sound, social signals  based on personal interests, time, location and activity. Early examples of this are already developed, such as the German Ambient News project funded by the Google News Initative [2]. There will be many more opportunities for story discovery with automated analysis of large datasets used as raw material for interesting stories: first draft, alarms, updates. As always, development comes down to money and it is reason to worry about the slow development in the field. Basic use of news automation, mainly creating stories about football, stock prices and real estate, is growing quickly, but more sophisticated features are not there. Besides financial resources, another problem is the general lack of an innovation culture in media. Unfortunately, doing product development and process innovation with money from Google is

not a substitute for serious research and development [2] even though Googles representative claim, We are all in this together [3].

## 6 Background

This paper is adapted from a talk that Adjunct Professor (Docent) Carl-Gustav Lindén, University of Helsinki (Swedish School of Social Science) and Affiliated Lecturer at the Sdertrn University, gave at the Norwegian Big Data Symposium 2018. There he shared his experiences of introducing news automation to journalists. He was the manager of the R&D project Immersive Automation (www.immersiveautomation.com). The primary aim with the project was to create a roadmap and a demonstration of a future news ecosystem based on automated storytelling and intense audience engagement paired with the belief that stories powered by data and machine learning will lead to a dramatically more personal and customized news experience with localisation of content as a key feature.

## 7 Acknowledgement

This paper is supported by European Unions Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

## 8 Disclaimer clause

"The results of this [publication] reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains".

## References

1. Nicholas D Allen, John R Templon, Patrick Summerhays McNally, Larry Birnbaum, and Kristian J Hammond. Statsmonkey: A data-driven sports narrative writer. In *AAAI Fall Symposium: Computational Models of Narrative*, volume 2, 2010.
2. M. Burkhardt. Ambient news lessons learned. https://medium.com/datenfreunde/https-medium-com-datenfreunde-ambient-news-lessons-learned-f190a48d8102, 2018. [Online; accessed 9-December-2018].
3. Madhav Chinnappa. We are all in this together. *British Journalism Review*, 28(3):50–55, 2017.
4. Mark Deuze and Mirjam Prenger. *Making Media: production, Practices, and Professions*. Amsterdam University Press, 2019.
5. M FRAYN. The tin men (london: Collins). also pp. 191-194 in stanley cohen and jock young (eds.) the manufacture of news. beverly hills, 1965.

6. Eli Goldberg. Fog: Synthesizing forecast text directly from weather maps. In *Artificial Intelligence for Applications, 1993. Proceedings., Ninth Conference on*, pages 156–162. IEEE, 1993.
7. Tony Harcup and Deirdre ONeill. What is news? news values revisited (again). *Journalism Studies*, 18(12):1470–1488, 2017.
8. Jaemin Jung, Haeyeop Song, Youngju Kim, Hyunsuk Im, and Sewook Oh. Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior*, 71:291–298, 2017.
9. L. Kung. Going digital how are legacy leaders transforming strategy, leadership and culture? Keynote at the Mediapiv conference, 2018.
10. Latar Noam Lemelshtrich. *Robot Journalism: Can Human Journalism Survive?* World Scientific, 2018.
11. Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197, 2017.
12. Leo Leppänen, Myriam Munezero, Stefanie Sirén-Heikel, Mark Granroth-Wilding, and Hannu Toivonen. Finding and expressing news from structured data. In *Proceedings of the 21st International Academic Mindtrek Conference*, pages 174–183. ACM, 2017.
13. Carl-Gustav Linden. Decades of automation in the newsroom: Why are there still so many jobs in journalism? *Digital Journalism*, 5(2):123–140, 2017.
14. Carl-Gustav Linden. Robot journalism. http://datadrivenjournalism.net/news_and_analysis/robot_journalism_the_damage_done_by_a_metaphor, 2017. [Online; accessed 16-February-2018].
15. S. Myles. How can we make algorithmic news more transparent? Paper presented at the conference Algorithms, Automation, and News, 2018.
16. E. Reiter. Where is nlg most successful commercially? https://ehudreiter.com/2018/10/30/most-successful-commercial-nlg, 2018. [Online; accessed 9-November-2018].
17. Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.
18. Dietmar Rösner. The automated news agency: Semtexa text generator for german. In *Natural Language Generation*, pages 133–148. Springer, 1987.
19. Andrew D Selbst and Solon Barocas. The intuitive appeal of explainable machines. 2018.
20. Jane B Singer. Who are these guys? the online challenge to the notion of journalistic professionalism. *Journalism*, 4(2):139–163, 2003.
21. J. Stray. The age of the cyborg. https://www.cjr.org/analysis/cyborg_virtual_reality_reuters_tracer.php, 2016. [Online; accessed 9-November-2018].