Department of Computer Science
Series of Publications A
Report A-2023-3

Methods for Automated Generation of Natural-Language Reports

Leo Leppänen

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki in Auditorium A110, Chemicum Building, on 21 April 2023 at 13 o'clock.

University of Helsinki
Finland

Supervisor
Hannu Toivonen, University of Helsinki, Finland

Pre-examiners
Albert Gatt, Utrecht University, Netherlands
Filip Ginter, University of Turku, Finland

Opponent
Ehud Reiter, University of Aberdeen, United Kingdom

Custos
Hannu Toivonen, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-9032-1 (paperback)
ISBN 978-951-51-9033-8 (PDF)
Helsinki 2023
Unigrafia

Methods for Automated Generation of Natural-Language Reports

Leo Leppänen

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
leo.leppanen@helsinki.fi

PhD Thesis, Series of Publications A, Report A-2023-3
Helsinki, April 2023, 72 + 73 pages
ISSN 1238-8645 (print)
ISSN 2814-4031 (online)
ISBN 978-951-51-9032-1 (paperback)
ISBN 978-951-51-9033-8 (PDF)

Abstract

The use of computer software to automatically produce natural language texts expressing factual content is of interest to practitioners of multiple fields, ranging from journalists to researchers to educators. This thesis studies natural language report generation from structured data for the purposes of journalism. The topic is approached from three directions.

First, we approach the problem from the perspective of analysing what requirements the journalistic domain imposes on the software, and how software might be architectured to account for the requirements. This includes identifying the key domain norms (such as the 'objectivity norm') and business requirements (such as system transferability) and mapping them to software requirements. Based on the identified requirements, we then describe how a modular data-to-text approach to natural language generation can be implemented in the specific context of hard news reporting.

Second, we investigate how the highly domain-specific natural language generation subtask of *document planning* – deciding what information is to be included in an automatically produced text, and in what order – might be conducted in a less domain-specific manner. To this end, we describe an approach to operationalizing the complex concept of 'newsworthiness' in a manner where a natural language generation system can employ it.

We also present a broadly applicable baseline method for structuring the content in a data-to-text setting without explicit domain knowledge.

Third, we discuss how bias in text generation systems is perceived by key stakeholders, and whether those perceptions align with the reality of news automation. This discussion includes identifying how automated systems might exhibit bias and how the biases might be – potentially unconsciously – embedded in the systems. As a result, we conclude that common perceptions of automated journalism as fundamentally 'unbiased' are unfounded, and that beliefs about 'unbiased' automation might have the negative effect of further entrenching pre-existing biases in organizations or society.

Together, through these three avenues, the thesis sketches out a way towards more widespread use of news automation in newsrooms, taking into account the various ethical questions associated with the use of such systems.

Computing Reviews (2012) Categories and Subject Descriptors:

Computing methodologies → Artificial intelligence → Natural language processing → Natural language generation
Applied computing → Arts and humanities
Software and its engineering → Software system structures

General Terms:
Algorithms, Design, Experimentation, Human Factors

Additional Key Words and Phrases:
natural language generation, automated journalism, news automation, report generation, multilingual text generation, artificial intelligence

Acknowledgements

Study hard what interests you the most
in the most undisciplined, irreverent
and original manner possible.

— Richard Feynman

Oh what a long and winding path it has been, full of ups and downs. This thesis is a culmination of a journey a decade in the making, and there are not enough pages in this manuscript to properly thank all the people who deserve my thanks for their support, friendship, and guidance along the way; you are all in my heart.

My sincerest thanks to Professor Hannu Toivonen, and the rest of the Discovery Research Group, for all the help, guidance and fun along the way; to Matti Luukkainen, for setting a 1st year humanities student on a path of computer science; to Arto Hellas (née Vihavainen), Petri Ihantola, Juho Leinonen, and the RAGE Research Group for everything in those first years; to Stefanie Sirén-Heikel, Hanna Tuulonen, Carl-Gustav Lindén and Lauri Haapanen, for being both excellent cross-disclipinary collaborators and the most patient explainers of what must have been elementary to you; to Associate Professor Albert Gatt and Professor Filip Ginter, the pre-examiners of this thesis, for their comments; to Professor Ehud Reiter, for agreeing to be my opponent; to "some NLG people", for you know who you are; to everyone involved with the EMBEDDIA, NewsEye and Immersive Automation research projects, for all the collaborations and support; to DoCS and HIIT for their funding and other support that made this thesis possible; and everyone else who has been a friend and a colleague.

And most importantly, the greatest thanks to Tiina, for her infinite patience and love; to Jaana and Ulla, for their support and understanding; and to Juha, for always believing in me.

Helsinki, March 2023
Leo Leppänen

Contents

List of publications

This thesis consists of the following publications, which are reprinted at the end of the thesis. In the introductory part of the thesis, the publications are referenced as Papers I-V. The author's personal contributions to each publication are listed below.

A Software Architecture for News Automation

Paper I: 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.* Association for Computational Linguistics, 2017, pp. 188-197, DOI http://dx.doi.org/10.18653/v1/W17-3528.

Contribution: The author of this thesis was responsible for a significant majority of the programming work underlying the paper, and was the major contributor to the text.

Paper II: Lauri Haapanen and Leo Leppänen, "Recycling a genre for news automation: The production of Valtteri the Election Bot," In *AILA Review*, Volume 33, Issue 1. John Benjamins Publishing Company, 2020, p. 67-85, DOI https://doi.org/10.1075/aila.00030.haa.

Contribution: The author of this thesis led the effort to create the underlying software system being described in the paper. Work on the paper, including discussions, analysis and writing, was equally shared between the two authors.

Planning and Structuring Documents for News Automation

Paper III: Leo Leppänen, Myriam Munezero, Stefanie Sirén-Heikel, Mark Granroth-Wilding and Hannu Toivonen, "Finding and Expressing News From Structured Data," In *AcademicMindtrek '17: Proceedings of the 21st International Academic Mindtrek Conference*. Association for Computing Machinery, 2017, pp. 174-183, DOI https://doi.org/10.1145/3131085.3131112.

Contribution: The author of this thesis took a leading role in the programming work described in the paper and contributed significantly to the writing of the article. The InterQuartile Range method of computing statistical outlierness was identified by Professor Hannu Toivonen, with further refinements to the broader method contributed by all authors.

Paper IV: Leo Leppänen and Hannu Toivonen, "A Baseline Document Planning Method for Automated Journalism," In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping University Electronic Press, Sweden, 2021, pp. 101-111, URL https://aclanthology.org/2021.nodalida-main.11.

Contribution: The author of this thesis ideated and developed the method being described, with feedback and suggestions from Professor Hannu Toivonen. The author of this thesis led the writing process, writing the bulk of the paper.

Ethical Considerations for News Automation

Paper V: Leo Leppänen, Hanna Tuulonen and Stefanie Sirén-Heikel, "Automated Journalism as a Source of and a Diagnostic Device for Bias in Reporting," In *Media and Communication*, Volume 8, Issue 3. Cogitation Press, 2020, pp. 39-49, DOI https://doi.org/10.17645/mac.v8i3.3022.

Contribution: The author of this thesis ideated and organized the paper and led the writing process, writing a majority of the content.

Chapter 1

Introduction

Information is a one of "two basic currencies of organic and social systems" [99, p. 445], with the search for, and use of, information being a "common and essential human behavior," "basic to human existence" [14, p. 4]. While technological advancement has made it ever-easier to produce data or raw information, our tools for turning that information into actionable knowledge have not necessarily followed suite. Streams of data have turned into fire hoses, with humans struggling to extract all that is extractible.

1.1 Automated generation of natural-language reports

This thesis is concerned with the automated creation of natural-language reports that express factual content. This task is studied in a dual context. First, technically speaking, the computational methods developed in this study fall into the field of Natural Language Generation (NLG) [35]. Second, the methods are applied primarily in the context of *news automation* [58, 59, 60, 100]. The primary interest of this thesis is determining *how to generate natural language reports from structured data for domains that emphasize factuality of content, using methods that are widely applicable.*

This broad question is approached through three distinct themes, capturing different aspects of the report generation process and how it aligns with perceptions of the technology's users and their core values.

Papers I and II contribute to the first theme, interested in how report generation systems should be architectured and constructed to be widely applicable within the journalistic field while taking into account journalism-specific concerns. This involves *identifying key requirements imposed on a*

report generation system by the journalistic context, and designing a high-level system architecture that matches those requirements. This work draws from previous works on the natural language generation architectures [e.g. 83, 89].

A second theme, contributed to by Papers III and IV, concerns a key technical step of the natural language generation process: deciding what information to include in the report, and how to structure said information. More specifically, the theme is concerned with *identifying broadly applicable methods for planning and structuring documents for news automation.* This work analyses and operationalises key results from journalism studies on how human journalists determine the "newsworthiness" of a piece of information [34] and how that information is then structured into textual narratives [106].

The third and final theme (Paper V) approaches the technology from a perspective of bias. It considers how newsroom professionals view the potential for bias in automatically produced news; whether their beliefs are well-founded; and what the consequences of potentially unfounded beliefs are.

1.2 Natural language report generation

The motivation for generating natural language reports stems, ultimately, from three things. First, there is an increasing need – or at least potential – for natural language generation systems to concretely aid human experts in their struggle to make sense of ever increasing torrents of information and data.

Second, there is an academic need to understand how to automatically mimic and approximate the types of complex mental processes undertaken by humans when they produce (and consume) natural language reports. How can we automatically detect that a piece of information in some way stands out from other pieces of information; how can we automatically organize all those data to produce coherent narratives, and how can we best translate those narratives into natural language?

The third aspect, which is intimately linked to the second but not a focus of this thesis, is that attempts to model a phenomenon allow us to better understand it. By creating a model of some process (e.g. structuring news texts), and then comparing the actions predicted by that model to those undertaken by the humans we are attempting to mimic allows us to compare the results of the two, facilitating a broader and more detailed understanding of the underlying phenomenon.

In terms of the first point (helping humans) in specific, to have a human journalist wade through the latest updated statistics from a national statistical agency is a non-trivial cost to even the largest newsrooms. Enormous amounts of news stories are not written because they are not economical to write given their potential readership. If automation brings the amortized cost of a single news text to near-zero, and those texts can be targeted at relevant audiences, it becomes financially feasible to produce news content with potential readership measured in single digits per text.

This type of a "long tail" can be significant. For example, by early 2000s some 38% of Amazon's book sales were of "niche" books not typically carried by brick-and-mortar stores [10, 11]. The ability to target this long tail is currently limited by the available human capital: assigning a highly trained expert journalist to spend an hour writing a news story that will only attract a handful of readers is unlikely to be a good use of journalistic resources. And even if that story "broke even," the opportunity costs of not having the journalist work on something with a greater readership (or rather, return on investment) are likely to be significant. The possibilities provided by automation in these contexts are extremely enticing. Indeed, visions of automation and AI as helpers of humans have been a key theme when the technologies have been discussed in media [19].

These limitations of the human capital are not limited to journalists or news data. By digitizing and making available larger and larger collections of historical newspapers, libraries around the world are providing researchers with a treasure trove of source material. However, the cost of understanding large corpora as a whole is very high: a human can read only so fast. Settings like these provide a potential ground for natural language generation systems that, together with state-of-the-art natural language *analysis* models, would allow human researchers to quickly obtain insights from even the largest collections of text. As part of the work leading to this thesis, we in fact demonstrated just such as system for analysis of historical newspaper archive material [77].

Nor are the human limitations only relevant in journalistic or scientific inquiries. The democratization of education through efforts such as massive open online courses (MOOCs) was already underway when the COVID-19 pandemic hit the world in late 2019 and early 2020. With the pandemic, the educational world was at once thrown into an online-first teaching scenario. Education suddenly becoming online-first, with teaching tied to neither a physical space or a specific time, raises questions about why courses should not be opened to even more students. In addition to institutional inertia, one important reason is the human resource cost of providing students

with feedback. This, too, presents fruitful ground for natural language generation, as we have shown in a related pilot study [55].

All of these problems – producing news content, understanding large datasets, providing students with feedback – can be eased with the use of natural language generation. In addition, they share a selection of key requirements, discussed in more detail in Section 3.1 from the specific perspective of news automation. This sharing of requirements highlights how all these different problems are fundamentally about *reporting*, using the term in a non-journalistic meaning. In this thesis, we approach the topic primarily from the perspective of producing journalistic texts – a process we refer to as *news automation* – but the methods described in this thesis are more widely applicable.

Automated systems producing natural language reports are not novel in themselves: Some of the very earliest natural language generation systems reported on stock market data [53, 54] and weather [7, 37]. Experience from the news industry, however, indicates that while there is significant *interest* in natural language generation systems [31, 58, 100], the market penetration of such systems remains limited [25, 40, 58].

Report generation approaches classically proposed in academic literature either fail to acknowledge the requirements of the journalistic domain (e.g. correctness) or must be extensively tailored to specific text types. A requirement for tailoring makes them too expensive to implement outside of a few key areas, such as weather and finance, while ignoring key domain requirements makes them fundamentally unfit for journalistic use.

1.3 Research questions

This thesis is organized around a singular primary research question: *how to best conduct data-to-text natural language generation for factuality-emphasizing domains, such as journalistic reporting.* This main research question is answered using three *research objectives*, listed below.

Research Objective I: Identify (a) requirements and (b) a high-level architecture for news automation

> This Research Objective entails an analysis of key values and conventions of journalism as both a practice and a genre of text. The identified key factors need to then be mapped into a set of requirements for journalistic NLG systems. This includes evaluating how the various NLG methods are suited to journal-

istic NLG. Finally, these requirements are used as a basis for developing an architecture for journalistic NLG.

Research Objective II: Identify broadly applicable methods for planning and structuring documents for news automation

Document planning is one of the most domain-specific aspects of NLG. This Research Objective involves the development of broadly applicable approaches to (a) identify newsworthy data points from large data tables and (b) produce coherent document plans from the identified data points. The methods developed must also align with the journalistic requirements identified in Research Objective I.

Research Objective III: Identify (a) how and why news automation systems can be biased, and (b) what could explain a hesitancy to accept news automation's potential for bias

This Research Objective involves identifying how news automation systems can exhibit bias and how the biases might be – potentially unconsciously – embedded in new automation systems. We also consider what might explain some key stakeholders' beliefs that news automation systems are unbiased.

1.4 Contributions of the thesis

The scientific contribution of this thesis is contained in the five original research publications listed as Papers I-V. The following overview of the publications includes their main contributions and how said contributions align with the research objectives identified above.

Paper I: Data-Driven News Generation for Automated Journalism

This paper analyzes the requirements for news automation, describes an architecture for news automation that fulfills said requirements, and presents a case study implementing the architecture. The paper contributes to Research Objective I, which is covered in Chapter 3.

Paper II: Recycling a genre for news automation: The production of Valtteri the Election Bot

This paper describes the identification, sense-making and fitting of genre conventions from the human-written texts into the automatically produced texts through the construction of the NLG software, contributing to the requirement analysis process. The paper also discusses the genre of the resulting news texts, and how it could be helpful to conceptualize them as exemplars of a new genre of journalistic texts. The paper contributes to Research Objective I, covered in Chapter 3.

Paper III: Finding and Expressing News From Structured Data

This paper analyses relevant journalistic theory on news values and newsworthiness. Based on that analysis, it describes a computational model for determining the newsworthiness of data points in structured data. The paper contributes to Research Objective II, which is covered in Chapter 4.

Paper IV: A Baseline Document Planning Method for Automated Journalism

This paper analyses relevant linguistic theory on the structure of news texts, and operationalizes that theory as a widely applicable baseline method for structuring fact-heavy news texts. The paper contributes to Research Objective II, which is covered in Chapter 4.

Paper V: Automated Journalism as a Source of and Diagnostic Device for Bias in Reporting

This paper investigates how perceptions of bias and objectivity are associated with journalistic NLG software and how those perceptions align with the reality of said software products. The paper contributes to Research Objective III, which is covered in Chapter 6.

1.5 Structure of the thesis

This thesis consists of five original research publications (Papers I-V) preceded by an introductory part with eight chapters (Chapters 1-8).

The next chapter (Chapter 2) gives a brief introduction to the field of natural language generation and establishes the primary context of this thesis, news automation.

Chapter 3 analyses requirements for a technical NLG approach for newsroom usage and describes a high-level architecture for an NLG system producing journalistic texts, answering Research Objective I.

Chapter 4 describes the document planning aspect of news automation, answering Research Objective II. It presents a method for identifying newsworthy data points from structured tabular data, as well as a method for structuring those data points into coherent documents without explicit domain knowledge.

Chapter 5 first discusses how natural language generation systems and methods such as those described in Chapters 3 and 4 should be evaluated. This discussion is then followed by evaluations of the works described in Chapters 3 and 4 respectively. The chapter finishes with a brief discussion on the general fitness-for-purpose of the described methods.

Chapter 6 discusses two key ethical concerns relating to news automation: bias and authorship. First, it describes biases that might be expressed in news automation systems, whether key stakeholders correctly identify risk of bias, and what consequences might result from any unfounded beliefs, covering Research Objective III. It then describes related works discussing how various parties interpret the authorship of texts written by news automation in terms of both the credit and the responsibility associated with authorship.

Chapter 7 starts by discussing how news automation methods could be further integrated into newsrooms. This is followed by a discussion on the limitations of the works described herein.

This introductory part is concluded by Chapter 8, which revisits the research objectives and provides an overview of some interesting avenues of future work. This concluding chapter is followed by the original scientific publications on which this introductory part draws.

Chapter 2

Background

This chapter provides an overview of the key concepts used in this thesis. First, we overview the natural language generation (NLG) problem and how it has been approached in previous works. Next, we give an overview of how automation and computational methods have been previously employed in newsrooms. Together, these sections give an overview of both the technological and usage contexts of this thesis.

2.1 Natural language generation

Natural language generation is a field of artificial intelligence research studying how to automatically produce textual documents in a natural language, such as English or Finnish [35]. As with any research field, there is some disagreement about what the exact boundaries of NLG are, but there is some agreement that the core problem of the field is generating text from *non-linguistic* inputs [20, 88].

At the same time, NLG can be understood more broadly as any process resulting in the generation of natural language, including from linguistic inputs [36]. Tasks included in this broader definition include summarization and translation of documents [35].

The work described in this thesis belongs to a specific subfield of NLG research: *data-to-text generation*. In data-to-text generation, the inputs of an NLG system are some type of structured data. This is in contrast to NLG systems that produce text from inputs such as images or video. This thesis uses "natural language generation" (and its acronym, NLG) as shorthand for data-to-text natural language generation.

This thesis is also closely related to a subfield of data-to-text generation known as *table-to-text generation*. Here, the system input is explicitly

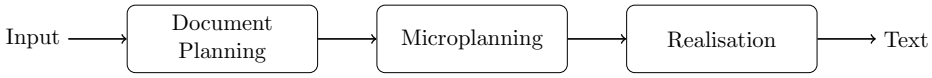

Figure 2.1: A very high-level view of the natural language generation subprocesses.

defined as a structured table of data [74]. The exact definition of a "table," however, is not strictly constant across various previously published works. While some use the term to refer to multi-dimensional tabular data [74], others' data is essentially key-value pairs [62]. Further, some works that explicitly use two-dimensional tabular data as input are described only using the more general data-to-text term [79, 80].

Due to the unclear nature of associated terminology, we have elected to not use the terms "table-to-text" for the work described in this thesis. This is further prudent in light of the nature of some of the contributions of this thesis: while most of the works are described using case studies that do employ tabular data, many aspects of the methods described in this thesis do not assume strictly tabular data.

2.1.1 Natural language generation subprocesses

As research into NLG goes back several decades, the field has seen a large array of different technical methods. In general, however, it is useful to conceptualize the larger process of "generating natural language text" as consisting of several sub-processes (Figure 2.1). Different works use slightly differing terminology and levels of detail in describing what the various subprocesses are, but on a very broad level it is common to differentiate between *document planning*, *microplanning* and *realisation* on a high level [35, 88, 89]. Some works also prefix these steps with distinct data analysis and interpretation stages [83]. In addition, the three main high-level processes can be separated further into more fine-grained processes [35, 89]. One such separation, the one used in this thesis, is shown in Figure 2.2.

Document planning

Document planning refers to selecting what pieces of information, called *messages*, are to be included in the final output, and how the document should be structured on a higher level [89]. Along these lines, document planning is often split further into *content determination* and *content structuring* subprocesses [89], the latter of which is also known as *text structuring* [35]. Notably, this formulation of the generation process assumes that

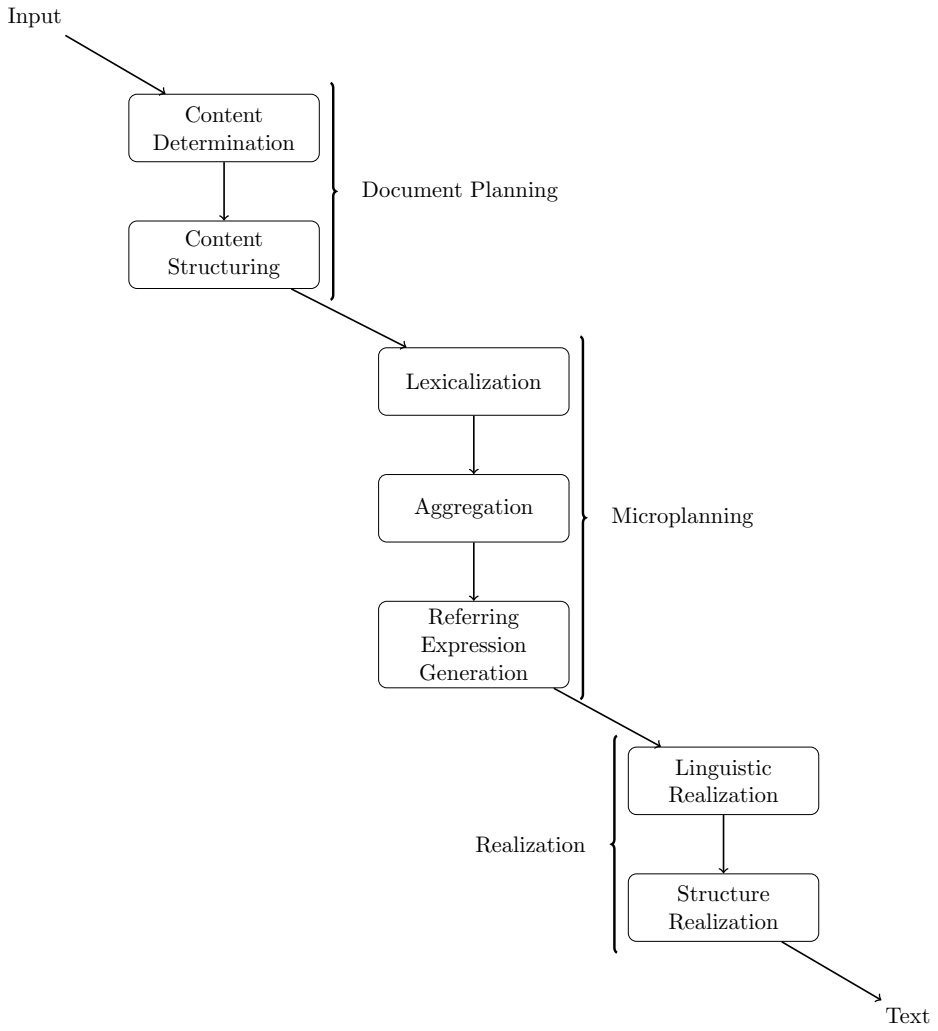

Figure 2.2: A more detailed view of the natural language generation sub-processes.

all the information that could be included in the text is already available at the onset of the document planning process. In some works, especially in the context of data-to-text generation, the document planning step is thus preceded by data (or "signal") analysis and interpretation stages, which produce more complex messages that might not exist directly in the underlying data, but can be identified from it [83].

These processes tend to be largely non-linguistic but very domain-specific [35]. Indeed, while various general approaches have been proposed for the later generation stages, document planning has traditionally been conducted through either hand-engineered approaches that are very domain-specific, or with machine learning approaches that are dependent on domain-specific datasets.

Microplanning

The document planning process is followed by a *microplanning* stage, which decides what linguistic expressions are to be used to convey the document planned during document planning [35, 89]. This stage, too, is often conceptualized as various subprocesses, most notably *lexicalization*, selecting of what words to use; *(sentence) aggregation*, grouping messages to sentences; and *referring expression generation* (REG), which decides how various *domain entities*, such as people or locations, are to be referred to [35]. For example, the referring expression generation stage would decide whether a person should be referred to using a full name and title ("Prime Minister Sir Winston Leonard Spencer Churchill, KG, OM, CH, TD, DL, FRS, RA"); a surname ("Churchill"); a pronoun ("he", "him", "they", "them") or some altogether different expression ("the third man from the left on the second row of the photograph").

Whereas the preceding document planning processes are largely domain-specific and non-linguistic, the processes of the microplanning stages are increasingly domain-independent and language-specific. Deciding what words to use is inherently linguistic, and while *some* domain-specific processing is required (for example, consider the difference in conducting REG for a Twitter bot and a system producing physical invitation letters for a formal ball: the first would likely use significantly less formal and terse language than the second), the degree of domain-dependence is clearly less than in document planning. Indeed, generic algorithms have been proposed for many of the subprocesses in this area.

Realization

The microplanning stage is followed by a process called *realization*, which can be conceptually split into *linguistic realization* and *structure realization* [89]. Here, linguistic realization concerns processes such as inflecting the words into their correct surface forms, while structure realization concerns processes such as adding the required markup symbols to display the document correctly. An example of the latter would be the insertion of HyperText Markup Language (HTML) tags to indicate paragraph breaks or that a certain span of text is a heading that should be rendered in a different font and size.

Of these processes, linguistic realization is largely domain-independent and highly general models and algorithms are plentiful in the literature. Structure realization is almost completely language-independent and largely domain-independent as well. Naturally, both processes need to account for some domain conventions, such as what style or register of language is to be used [1, 16].

2.1.2 Classifying NLG systems

Historically, NLG systems have applied different task-specific algorithms and hand-engineered grammars or templates for the generation. More recently, academic interest has shifted towards machine learning methods, most significantly neural networks.

These neural network-based approaches often promise to produce more fluent and varied language while demanding less engineering effort, as the systems learn from human-written texts [28]. However, they suffer from several key problems that limit their practical applicability. One critique relates to both the existence and cost of producing and cleaning suitable training data. Another problematic aspect relating to the present neural state-of-the-art is the tendency of neural systems to *hallucinate* content, i.e. produce text that is not grounded in the system inputs [28, 51]. The degree to which these criticisms are valid naturally depends on the context where the NLG system is expected to function. As a solution, some authors have suggested combining neural models with symbolic logic [68].

NLG systems can also be classified based on their architectures. Systems reported in previous works form a spectrum based on whether their designs follow the aforementioned conceptual decomposition or not. Approaches forgoing any decomposition have been especially common in the case of approaches based on neural networks [27, 28, 35]. More recently, even neural models have re-introduced different decompositions, even if the

systems are still commonly trained using an "end-to-end" approach [79]. Some works have even employed a "classical" pipeline architecture built from neural components [32].

2.2 Automation in newsrooms

Newsrooms are not new to disruptive technologies. In a way, the printed newspaper was *born* from such a disruptive technology, the printing press [5]. Yet, despite being born out of what could be called the "second information revolution" [99], the news industry has not always embraced all technological advances. Hesitancy towards automation is well demonstrated by how the automation of printing was met with strikes from as early on as 1963 [49, p. 344]. To this day, fears of jobs being lost to automation remain an "omnipresent" concern in discussions regarding news automation [40]. Despite this – at least historical – skepticism, the outputs of news media have presented automation and AI in more optimistic than negative terms [19].

2.2.1 News automation

The use of computers to generate news stories goes back at least to the mid 1980s, with some of the earliest NLG systems reported in academic literature being a weather reporting system [7], a system for producing news texts about terrorist attacks [21] and a generator for stock reports [53, 54].

Despite the aforementioned early news text generation systems, use of natural language generation methods in real-world newsrooms remained limited until the 2000s. Early examples of large-scale real-world automated news text generation include Statsheet, which began to publish automatically generated college basketball coverage in 2010, and Thomson Reuters which announced in 2006 that it would automatically produce financial news coverage [105]. Other early examples of text generation for journalistic purposes include *Los Angeles Times*, which produced automatic earthquake coverage by 2014 [12]. By 2017, many news agencies across Europe either had adopter, had trialed, or were developing methods for automated generation of news text, primarily in the domains of sports and finance [31]. The interest has not waned since [100].

While it is clear that the use of natural language generation methods in the newsrooms is of interest to the news industry, the academic discussion around these methods and their effects remains complicated. First, systems that automatically produce textual content for journalistic purposes have been described in the academic literature using widely varying terminology. The most commonly applied terms include "robot journalism", "al-

gorithmic journalism", "automated journalism", "machine-written news", "computational journalism" and "news automation" [58, 60].

Second, much of the relevant academic literature uses these terms in slightly differing ways. For example, Dörr [26] defines "algorithmic journalism" as the application of natural language generation techniques to journalism. On the other hand, Graefe [39] defines "automated journalism" as the production of news stories without human intervention. The important distinction between these two definitions lies in what drives the definition: Dörr [26] defines the term through the underlying technological process, while Graefe [39] bases their definition on how the technology integrates into the newsroom. By defining the term through the technology, Dörr [26] also imports the discussion surrounding the input of NLG, namely that rephrasing, summarization or translation are excluded from the definition (See Section 2.1).

In terms of how the automation integrates into the newsroom, both Dörr [26] and Graefe [39] use language indicating that the produced texts would optimally be published as-is, with Dörr [26] stating "published mainly automatically" and Graefe [39] defining the final step of algorithmic news generation as publishing the story "either automatically or after editorial review". While others share the input aspect of Dörr's [26] technology-driven definition, many other definitions are less strict about the output of the process. For example, Dierickx [25] defines news automation as consisting of "transforming structured data into texts in natural language or other form of visual representation", while Haapanen [42] describes it as referring "to algorithmic processes that convert data into a user-friendly form." Notably, both of these definitions allow for "news automation" to produce graphs, figures and other non-text content.

In this thesis, we adopt the term "*news automation*" following the example of Lindén, Dierickx and others [25, 59, 60, 100]. We define the term as *automated production of natural language reports for journalistic purposes from structured data*. This definition balances both the technology-driven definition used by Dörr [26], as well as the integration-driven definition of Graefe [39], while allowing for the more complex outputs included in the definitions of Haapanen [42] and Diedrickx [25], such as the inclusion of various visualizations.

Finally, we note that the definition leaves purposely undefined who the audience of the produced report is, allowing for both direct (or limited-oversight) publishing to the news reader, as well as the newsroom-internal use of the reports as, for example, news alerts or "first drafts" of news texts that a human journalist would then further process. This type of workflow,

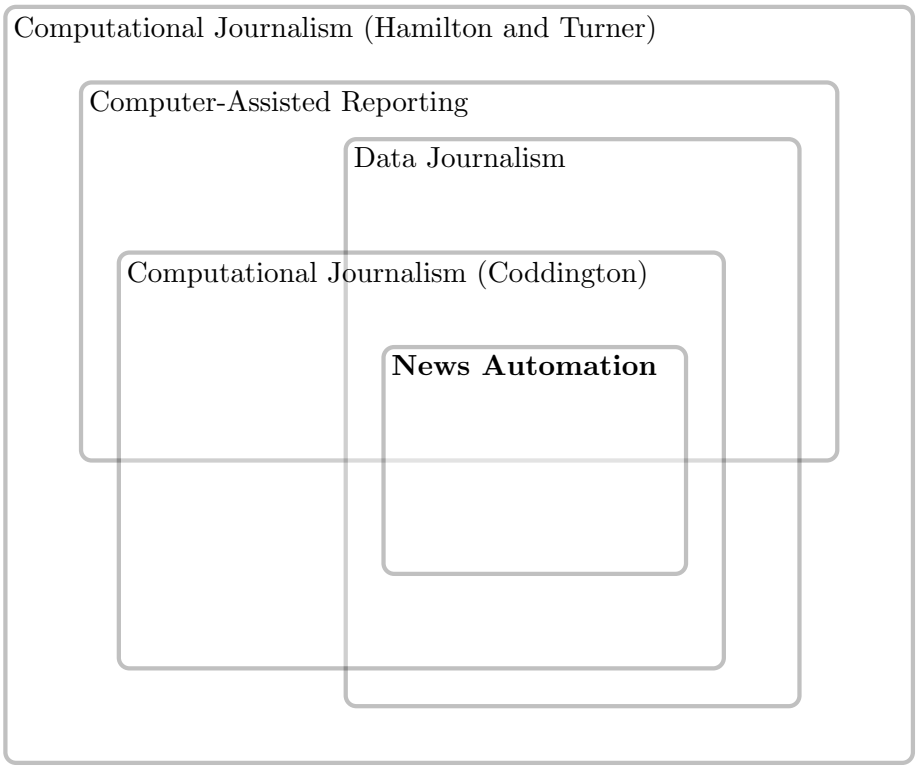

Figure 2.3: Mapping news automation into various related terms. Based on definitions as described by Coddington [18].

where automation highlights potential news stories, has been called *computational news discovery* in previous works [24]. Indeed, as discussed later, we believe this method of employment to be the most promising near-future use of news automation

2.2.2 Positioning news automation

The lack of terminological consensus – discussed above – also indicates that it is not necessarily obvious how the various automated methods integrate into the newsrooms. To better understand the definition of news automation as used in this thesis, we need to establish several related terms involving the use of automation and data-driven processes in the newsrooms (Figure 2.3).

According to Coddington [18], one of the early key terms is *Computer-Assisted Reporting* (CAR). While the exact definition of Computer-Assisted

Reporting has transformed over time, for example through the diffusion of the methods into "standard" journalistic work, it "includes techniques such as data searches on the web, spreadsheet and/or statistical analysis of various public records, and geographical and other information mapping" [41, p. 718]. By this very broad definition, it would include almost all practices of modern (investigative) journalism. In relation to news automation, Computer-Assisted Reporting describes a superset of news automation tasks: while news automation involves (or at the least *can* involve) e.g. statistical analysis, it is unlikely to incorporate general web searches.

Another related term identified by Coddington [18] is *data journalism*. Data journalism refers to a wider array of practices surrounding the use of (especially open) data, its analysis and visualization. While clearly related to Computer-Assisted Reporting, data journalism is less coupled with investigative journalism in favour of methods that "allow the public to analyze and draw understanding from data themselves, with the data journalist's role being to access and present the data on the public's behalf" [18]. Using this definition, data journalism is well-aligned with news automation. One way of describing news automation in this context would be to view it as a *method* of data journalism.

The final related term identified by Coddington [18] is *computational journalism*, which has been used with varying definitions. Some of these definitions are very broad, such as that of Hamilton and Turner [43] who define it as "the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism." Others are more strict, with Coddington [18] defining it as "technologically oriented journalism centered on the application of computing and computational thinking to the practices of information gathering, sense-making, and information presentation". Of these two definitions, Hamilton and Turner's [43] description is very broad and can be seen as encompassing everything discussed so far. Coddington's [18] definition, on the other hand, is more closely aligned with news automation, as we could describe news automation as one example of how the computation and computational thinking could be practically applied to the aforementioned tasks.

Against this backdrop, our definition of news automation emerges as a method of conducting data and computational journalism as defined by Coddington [18]. Furthermore, when used in the newsroom-internal operation mode, it can be viewed as a method for Computer-Assisted Reporting, where reporters inspect datasets using the reports produced using news automation methods. Viewed in this way, news automation is an augmentative technology, helping human journalists obtain insights into data

faster. On the other hand, news automation – especially when fully automated – could be viewed not as an augmentative technology, but as one replacing the human. Even if analyzed thus, it remains a component of data journalism, merely further removing the journalist (as the creator of the news automation software) from the audience.

Chapter 3

Generating reports for journalistic purposes

While newsrooms have expressed significant interest in the use of news automation methods, the market penetration of news automation has been surprisingly slow given that academic publications on the automated generation of news text go back to at least the 1980s [7, 21, 53, 54].

In this chapter, we first describe how the journalistic context influences various design decisions that have to be made when designing news automation systems. Then, we determine how these considerations can be mapped into a set of requirements for news automation systems. The results of these analyses are then reflected against the various NLG approaches described in previous works, and used to identify a high-level architecture for news automation. This discussion is based on Papers I and II. An evaluation of this approach is presented later, in Chapter 5.

3.1 Requirements for news automation

In order to identify requirements for news automation we view the software development process, including the requirements analysis process, as an example of "recontextualization" [61, p. 154]. Recontextualization is a process wherein conventions of one field and genre are extracted and fitted into another context. For example, the news domain is strongly associated with a norm for objectivity. During the news automation software development process, this norm must be applied – recontextualized – in the specific context of the software. On a broad level, we can view the creation of any NLG software as an act of recontextualizing some human-written genre into a computer-written equivalent.

The recontextualization process is fundamentally a sense-making practice [61, p. 155]. It involves first identifying the important aspects of the source domain (such as the objectivity norm, described below), and then determining how those aspects are best fitted into the automated text domain. This process results in a set of system requirements, such as described in Papers I and II and overviewed below.

When recontextualizing human-written news text into news automation, the relevant genre conventions span the whole generation process. That is, they go from surface level aspects such as the font and formality of the text to values that influence how certain attributes of the software are prioritized.

The insights of this chapter derive from our experiences in incorporating domain conventions in the context of producing a news automation system for election coverage (Papers I and II). As election news coverage presents a very prototypical news reporting context, the results should be extensible to a wide array of "hard" news.

Objectivity

A very high-level norm in (especially Western) journalism has classically[1] been the objectivity norm [95]. It has been defined as "reporting something called 'news' without commenting on it, slanting it, or shaping its formulation in any way" [95, p. 150]. Others investigate it in terms of components such as "detachment, nonpartisanship, a style of writing called the 'inverted pyramid', facticity, and balance" [66, p. 2].

This objectivity norm influences news automation systems most directly as a requirement for **accuracy**: even if all else fails, the objectivity norm demands that the output produced by the news automation system is factually correct. The norm also affects processing stages such as lexicalization, by demanding that the word choices made by the system are impartial.

The norm also dictates that news text content is driven by "newsworthiness" (irrespective of the vagueness of the concept) and structured using the inverted pyramid model. This means, broadly speaking, that the news text should introduce answers to the most important questions at the start of the text [78, 103]. Together, these conventions influence how the factual content of the document is selected and ordered.

[1] For a discussion on how this norm is being re-interpreted in a "post-truth" era, see e.g. [13, 64].

Fluency

In addition to being factually correct, any news automation system has a fundamental requirement for producing output that is understandable. If the system produces incoherent text, it is not fit for a journalistic purpose. As such, there exists a requirement for at least some minimal level of textual **fluency**.

At the same time, it is not obvious how fluent the output of the system *must* be. For example, one might attempt to derive a threshold from whether the reader understands the messages the system is attempting to convey. Yet texts of this minimal level of fluency are unlikely to be well-received by the general audiences. At the same time, it seems reasonable that a news automation system targeting *journalists* (e.g. one producing newsroom-internal "news alerts") is allowed significantly more leeway in how (non)fluent the produced texts are. In other words, the required level of textual fluency depends on both how the system is employed and who the target audience is.

Transparency

Another high-level norm in journalism is transparency. It requires that "people both inside and external to journalism are given a chance to monitor, check, criticize and even intervene in the journalistic process" [23, p. 455]. This norm can be mapped relatively easily into a news automation system by requiring that the system itself be **transparent**, or perhaps even **explainable**: For the actors both inside and external to the newsroom to monitor, check and criticize the way in which the news automation system functions, they must be able reason about what the system is doing and why. This property is further important for ethical reasons, as discussed in Chapter 6.

Modifiability and transferability

The news automation system must be one that allows for easy and targeted system **modification** to correct any mistakes the generation process is making. This requirement for modifiability stems from another facet of the transparency norm, namely a requirement that the journalistic process must be intervenable [23, p. 455].

The modifiability requirement can also be framed as a requirement for **transferability** between different news production subdomains. The software should be modifiable that it can be transformed into a system working

from a distinct dataset covering a different topic. Failure to fulfill this requirement has been identified as a weakness of many current and previous news automation systems [58].

Data availability and topicality

News automation systems derive value from input data. As such, any news automation system must consider both the *availability* and *topicality* of the data being ingested and the news being produced. Even a ground-breaking news automation system is worthless if it requires data that is either not available, or only becomes available so long after the fact that the "news" is no longer news.

Other considerations

In addition to the above high-level considerations, genre conventions influence every aspect of the resulting NLG software. They define how long the resulting text should be, what words and registers are to be used, and what the tone of the text should be. They even influence how the resulting text be rendered to the viewer: how long the lines are, how large the characters, what style the font.

3.2 Suitability of technologies for news automation

The above requirements highlight a weakness in the currently available NLG technology. As discussed in Section 2.1, recent NLG methods based on machine learning are yet to achieve sufficient maturity in terms of the textual output's faithfulness to the underlying knowledge base. This phenomenon is commonly known as "hallucination" [51] and presents a significant problem in terms of the objectivity norm.

A further problem with neural approaches to NLG is their opaque nature. Many machine learning approaches are commonly referred to as "black box systems", meaning that it is at least currently very difficult for humans to obtain a good understanding of why the systems make certain decisions [93]. This is problematic in terms of the requirement for transparency.

As a consequence of their complexity, neural systems are also extremely difficult to modify outside of retraining, potentially with a different corpus. This limits their maintainability, as fixing errors or making minor modifications requested by domains experts is potentially very difficult [86, 87].

As such, they are poorly aligned with the intervenability and modifiability requirements identified above.

Machine learning methods in general are also dependent on datasets that might not realistically exist for many domains, languages, or combinations thereof. Even when these corpora exist, they can be unaligned or incomplete in relation to the gold standard texts written by journalists [52]. As a consequence, while neural methods are highly transferable in theory, their *practical* transferability within any specific newsroom is not necessarily suitably high.

In light of these problems, it is not surprising that the commercial NLG landscape continues to be dominated by methods based on rules and templates [20]. At the same time, these methods are viewed by news industry professions as difficult or costly to transfer between news domains [58].

Overall, we interpret the above as suggesting that machine learning approaches present users with a low "quality floor". This means that the user cannot be confident that neural systems will always produce output of at least acceptable quality. Others have attributed this lack of robustness in neural methods to "the emphasis of academic evaluations on average-case performance instead of on worse-case performance" [85]. On the other hand, especially neural systems show some indication of producing more natural text than their rule-based counterparts [28]. In doing so, their best-case outputs might present a very high "quality ceiling". Rule-based systems, on the other hand, appear to have a higher quality floor: they do not suffer from the catastrophic errors caused by "hallucination". At the same time, they appear to have lower quality ceilings in terms of output fluency.

Given that both rule-based and neural approaches to NLG appear to suffer from distinct problems, it is natural to ask how this state of affairs might be improved. Overall, there appear to exist three avenues of improvement:

1. addressing and eliminating the downsides of rule-based systems;
2. addressing and eliminating the downsides of neural or otherwise machine-learning systems; or
3. producing hybrid systems that combine the upsides of rule-based and machine learning-based systems while avoiding the downsides of both.

In the following section, we present a software architecture for news automation that primarily targets the first avenue of improvement. However, through modular design, the identified approach also lends itself to integrating aspects of the third avenue of improvement.

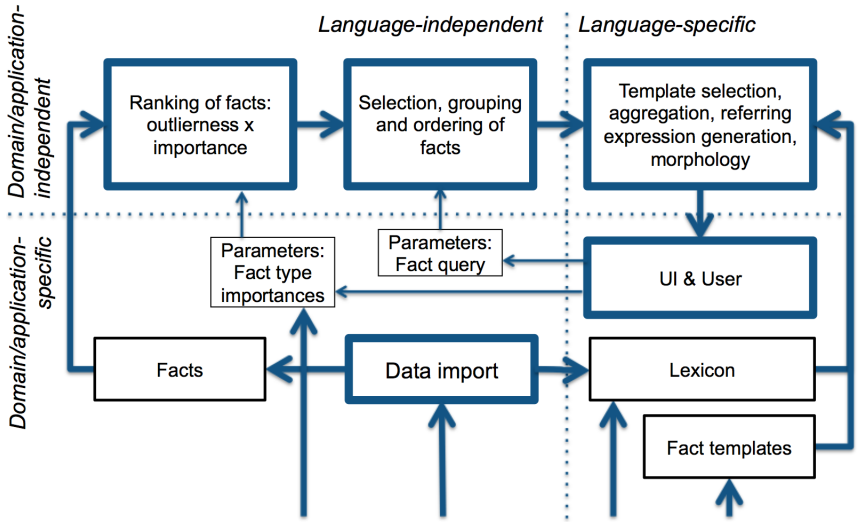

Figure 3.1: A high-level view of the architecture as presented in Paper I. Boxes with bold blue borders represent (clusters of) software components, while the boxes with thin black borders represent (meta)data used to parametrize the software components during generation.

3.3 A pipeline architecture for news automation

In Paper I we describe a news automation system architecture (Figure 3.1) built as a pipeline of separate components. As discussed earlier, the general approach of separating the various conceptual language generation stages into separate software components is not novel [35, 83, 88, 89]. However, the architecture described here tailors these previous high-level approaches to an architecture highly suited for fact-heavy journalism while maximizing transferability and modifiability. The architecture also considers multilinguality from the ground up.

The architecture presented in Paper I can be viewed as employing the *separation of concerns* design principle through modularity [75, 102]. As in many other non-neural generation systems, the broader generation process is split into conceptually distinct processing phases (See Section 2.1). Each of these phases is then associated with a separate software module or component, which correspond to the blue boxes in Figure 3.1. In the case study implementation, the "template section, aggregation..." box was decomposed into a sequence of separate components.

While the general modularization approach is common in previous natural language generation works [88, 89], we take this modularization fur-

ther by explicitly considering how to separate the subprocesses into general parameterizable software components and their domain-specific and/or language-specific parameters. This is demonstrated in Figure 3.1 by the separation of the various components and their parameters into the different quadrants.

To increase transferability, the architecture employs a simple meaning representation: atomic pieces of information are represented as a flat set of *facts*, consisting of a collection of key-value pairs. Each fact has a singular "value", which represents the core – usually numeric – information transferred by that fact. Additional key-value pairs then describe how to interpret that value or what the broader context of that value is.

In Paper I, we describe the core value as a `what` value, and the corresponding metadata describing how to interpret said value as the `what_type` value. Additional metadata fields include `who` and `where` fields, as well as their corresponding `who_type` and `where_type` fields, which again describe how to interpret the base values. For example, a fact might consist of the `what` value '`15`', and a `what_type` value '`nr_of_seats`' to convey the idea that some party obtained 15 seats in an election. Metadata key-value pairs `who=SDP` and `who_type=party` would then add the context that the number of seats belongs to the Social Democratic Party of Finland, and that said entity is indeed a political party. The above field-naming scheme is derived from the journalistic context, where it is often useful to think of "who did what and where." In later works, we have replaced the terminology with more generic terms (`value` rather than `what`, `location` rather than `where` etc.) but have retained the general format of representing atomic pieces of information.

A crucial part in the language generation process is that of turning the aforementioned atomic units of information into a coherent textual document. This involves selecting what pieces of information to include in the text – in the case of Paper I, selecting a handful of facts from among hundreds of thousands – as well as ordering the selected facts into paragraphs. In Figure 3.1, these steps correspond to the two component clusters in the top left quadrant: "ranking of facts" and "selection, grouping and ordering of facts".

A key problem in the development of the architecture was ensuring that content selection and structuring processes are minimally dependent on either the language of the text being generated or the domain of the text. This topic is discussed in detail in the next chapter, which overviews Papers III and IV.

The domain and language considerations are perhaps the most intermixed in the lexicalization stage, where the system decides what words and phrases ought to be used to express the meaning of the planned document. The architecture described in Paper I approaches this subtask through phrase-level templates defined using a custom templating language. While template-based approaches to NLG are not novel,[2] they provide several key benefits in the context of news automation.

First, as shown in Figure 3.1, their use allows the domain-specific "fact templates" to be separated from the template selection component, which thus becomes domain-independent.

Second, templates can encode lexicalization information in a format that is easily understood by domain experts in comparison to, for example, grammar-based approaches. While it is not reasonable to expect that a journalist with no technical or linguistic background would be able to contribute directly into a grammar-based news automation system, we observed in practice during the development of the system that journalists were able to contribute directly through the generation of phrase templates.

In our templating language, a simple template might be written as the template string "`{entity} won {value} seats in {location}`" which is then associated with a simple rule, such as "`value_type = nr_of_seats, value > 0`". Here, segments in {brackets} are *slots* and expose values from the facts the templates are associated with. Our templating language is further designed to support system multilinguality: each template string can be associated with a language code prefix (e.g. "`fi:`" for Finnish), and multiple template strings can be associated with a shared rule in a template group.

As a next step in the architecture, these templates are then aggregated into longer sentences. The goal of aggregation is to improve the fluency of the system output. For example, in the case study instantiation described in Paper I the phrases 'Party X got 2 seats' and 'Party X got 792 votes' would be aggregated as 'Party X got 2 seats and 792 votes'. In Paper I, this aggregation is done using an extremely simple prefix-matching approach. While this approach has the benefit of being domain-independent, its simplicity causes problems in terms of text quality and correctness, discussed in Chapter 5.

Following aggregation, the document still contains some unlexicalized content. For example, domain entities such as parties are still referred to by their internal identifiers. A Referring Expression Generation phase

[2] As demonstrated by early discussions regarding whether templates constitute "real" NLG [22].

then determines whether any specific reference to the entity can be e.g. a pronoun-like term (such as "the candidate"), a short phrase that is understandable once the entity has already been introduced (e.g. "Smith") or a complete reference best suited for first introducing the subject (e.g. "John Smith"). As shown in Figure 3.1, Paper I separates this task into domain-independent algorithm and a domain-specific lexicon.

Finally, the planned and lexicalized document must be realized into a natural language text. This includes inflecting various words to their correct morphological forms – achieved through the use of 3rd party tools and a short dictionary for uncommon terms such as some names – as well as the addition of HTML tags to enable the generated texts to be displayed as web pages. In Figure 3.1, we separate the domain-independent morphological realization phase ("morphology", in Figure 3.1) from the domain-specific rendering decisions ("UI" in Figure 3.1).

As described above, the architecture largely separates the language-specific aspects of the generation process from the domain-specific aspects. By separating the language specific from the domain specific, the resulting architecture can be more easily reused in a new domain (as the language-specific components can be reused), and is easier to make multilingual (as the language independent processes can be shared). An evaluation of the architecture is provided in Chapter 5.

Chapter 4

Planning journalistic documents

One of the key problems with practical news automation applications is the non-transferability of software [58]. In Section 2.1, it was identified that some of the most domain-specific aspects of natural language generation are the early processing steps associated with deciding what content, and in what order, is to be included in the document. These content selection and document planning processes are commonly handled either using hand-engineered methods that do not generalize, or through machine learning approaches that present a general *method*, which still depends on domain-specific data.

Papers III and IV present methods for both content selection and document planning. These methods are not based on machine learning, and thus do not require the existence of training data. At the same time, they are widely applicable within several subdomains of news text generation. Since the later parts of the NLG process tend to be significantly less domain-specific, reducing the domain-dependence on these critical subtasks causes a meaningful reduction in the domain-specificity of the NLG software in totality. This, in turn, addresses the aforementioned transferability problem, allowing the easier creation of news automation systems, which is a prerequisite for increased real-world use of news automation systems.

The next section introduces a method for identifying newsworthy messages from large datasets, i.e. conducting content selection. This work is published as Paper III. The following section then introduces a document planning method that integrates with the aforementioned content selection method. This work is published as Paper IV.

4.1 Identifying newsworthy data points

The creation of a news automation system entails the recontextualization of journalists' conventions into the news automation system (See Section 3.1). One of the key conventions in journalism is a concept of *newsworthiness*. This concept is innately tied to the idea that journalism includes a function of *gatekeeping*, or "winnowing down a larger number of potential messages to a few," [98, p. 100] where messages with high newsworthiness are passed on to the reader, while those with low newsworthiness are left out.

Newsworthiness is a complex phenomenon which incorporates broad societal and cultural value systems. It has been studied extensively outside of computer science, with Galtung and Ruge [34] providing one key characterization of the various component values that influence the likelihood that a piece of information is published. They include factors such as "absolute intensity", "scarsity", "reference to elite nations" and "reference to something negative." Later works [44] have built upon Galtung and Ruge's [34] work, adapting it for an evolving society.

In Paper III, we take these theoretical works and operationalize them in the context of a news automation system. This involves identifying which of the various aspects identified in the theoretical works translate meaningfully into computable metrics, how those metrics should behave in isolation, and how they should be combined into a single numeric value representing a single message's newsworthiness.

We group the various aspects described by Galtung and Ruge [34] and distill from them four distinct computable factors. The first factor, *topicality*, incorporates the observation that more recent events associated with other things in the current public discourse tend to be more newsworthy. Second, a factor of *outlierness* encodes the observation that events and things that are somehow surprising in the context of other events tend to be more newsworthy. Third, a factor we label *interestingness* incorporates widely held views that affect newsworthiness, such as that events associated with certain noteworthy or famous people or places tend to be more newsworthy than those associated with relatively unknown people or places. Finally, we include a *personalization* factors, which allows the encoding of beliefs similar to interestingness, but from the perspective of the individual reader, rather than broader societal tendencies.

This leads us to a theoretical formulation of newsworthiness of an event e, in the context of a public discourse d and other related events E_r as

$$N(e, d, E_r) = T(e, d) \times O(e, E_r) \times I(e) \times P(e) \tag{4.1}$$

where T, O, I and P are functions that assign numerical values to top-

icality, outlierness, interestingness and personalization, respectively. As noted, some of these decisions are contextual, incorporating information from other associated events and the broader public discourse.

Because of the difficulty of defining the public discourse d, and the uniform temporal nature of the data use in Paper III, we effectively omit the topicality aspect in that work. In other systems, we have employed simple formulations based solely on the recency of a data point, where older events get lower topicality scores.

For outlierness, Paper III describes a method based on interquartile ranges to produce a numeric value that encodes how different one data point is in the context of a sample of similar data points. More specifically, this "outlierness" value is minimal between the first and third quarter points of the sample, and increases further away from these central points.

For interestingness, the method relies on message metadata and hand-engineered weighing. This is implemented using multiplicative weights that state, for example, that messages related to municipality-level results are the most interesting in a municipal election context. These weights were implemented for the type of the location, the type of the entity, and the value type.

Finally, we incorporate the personalization aspect by allowing users to set the focus of a news text. For example, when a user indicates they are interested in a specific municipality, the document planner forbids the inclusion of information about other, unrelated, municipalities.

We model these four aspects as non-negative real numbers unbounded from the top. This allows us to multiply the four factors into a single non-negative real number representing an estimate of a message's newsworthiness. This estimate can then be used together with other considerations to plan the contents of a news article, discussed in the next section. The decision to bound the values as described stems from an observation that while the concept of "not interesting" is meaningful, the concept of "maximally interesting" is not well defined.

A related facet of our work in operationalizing newsworthiness in this manner is that it also provides a method for prodding human journalists' news values for biases. It should be possible to compare any decision made by our model of newsworthiness to the decisions made by human journalists in a similar situation. Any differences between our model and the human journalists' behaviour would then be attributable to either the inherent fuzziness of human mental processing, or to a discrepancy between the computational model and the humans' internal models of newsworthiness. We return to inspecting human biases in Paper V and Chapter 6.

4.2 Planning news reports

While newsworthiness plays a major role in how news texts are constructed, it is far from the only factor. In Paper IV, we present some of the relevant theory from studies of how human-written news texts are structured, and then build on that theory to present a broadly applicable method for producing coherent fact-heavy news texts.

Paper IV synthesizes three observations. The first is the well-known "inverted pyramid" structure of news [78, 103], stating that news texts tend to present the most newsworthy or important information first. Second, White's [106] orbital theory of news describes how the paragraphs of "hard news" text form an orbital two-tier structure: hard news stories tend to consist of a central key paragraph that is at the very top of the story, followed by "orbiting" supporting paragraphs that are largely reorderable. Finally, we observe that White's [106] theory closely mirrors the Rhetorical Structure Theory [63], only on a larger scale.

Based on these observations, we then describe a broadly applicable method for constructing and ordering topically coherent paragraphs. The fundamental assumptions of the approach are the existence of a numerical estimate of newsworthiness, provided here by the method described in the previous section, and metadata that allows us to estimate how semantically related two pieces of information (messages) are.

The coherence of the produced text is ensured by three factors: a *contextual similarity* factor, a *topical similarity* factor, as well as a *penalty term* discouraging unfocused narratives.

Contextual similarity captures the notion that a message about some location or time is likely to be followed by messages discussing the same location or time. The contextual similarity between two messages A and B is determined by inspecting whether the various metadata fields of the two messages share same values. Each field sharing the same value (i.e. in cases where both messages discuss, for example, the same country or instance of time) contributes a multiplicative weight to the contextual similarity factor. The similarity is specifically defined to be zero when there are no shared fields between the two messages.

The *topical similarity* term captures the notion that a message discussing some specific aspect (e.g. unemployment figures) is more likely to be followed by additional messages about the same topic, rather than about some altogether distinct topic. This is done by comparing the messages' `value_type` fields, which are analogous to the field `what_type` in Paper I.

Our method for calculating topical similarity assumes the fields contain hierarchical labels. For example, in the system described in Paper IV, a

message having `health:cost:hc2:eur_hab` as the contents of its `value_type` field indicates that the message's `value` represents the healthcare-related cost of rehabilative care measured as euros per inhabitant. Similarly, `health:cost:hc41:mio_eur` indicates contents of the `value` are to be interpreted as the healthcare-related cost of imaging services measured in millions of euros.

We describe two slightly different methods for determining the degree of similarity between `value_type` values. The first method inspects the length of the unshared suffix counted in colon-separated segments. For example, the above examples both share the prefix `health:cost` indicating they both discuss the same general topic, but concurrently have relatively long unshared suffixes. This method makes the implicit assumption that both labels are equally long. As such, we also present a slightly more complex variant that compares the length of the shared *prefix* normalized by the lengths of the labels, accounting for labels of potentially different lengths.

The *penalty term* – labeled "set penalty" in Paper IV – allows for inclusion of tangentially related messages, while ensuring that the text does not endlessly "drift" from the original topic. This allows, for example, a text discussing the economic situation of one country to make a comparison to another country. Our formulation assumes that the input messages are split into a *core set* of messages that are unambiguously related to the crux of the story being generated, as well as into an ancillary *extended set* of messages that could be included, but should be limited to a supporting role. In our case study, this split is based on geography and user input: if the system is asked for a story about France, all messages pertaining to France are allocated to the core set, while all messages about other countries are allocated into the extended set. When considering what message to add to an in-progress document plan, the extended set's messages' newsworthiness values are multiplied by a penalty term of $\frac{1}{d}$, where d is the distance to the preceding core message. This discourages long chains of extended messages while allowing short digressions.

These three factors are applied when considering with which message to extend a paragraph plan under construction. For each potential "candidate" message that could be the next message included in the paragraph, that "candidate" message's inherent newsworthiness value is multiplied by weights representing the above three factors. The maximally scoring candidate is then selected for inclusion in the document, after which the process repeats until either the end of the document or the paragraph is reached.

When beginning a completely new paragraph, we prohibit the paragraph from starting with an extended set message (i.e. a message about

something ancillary, such as a message about Germany in an article about France). We also require that the paragraph's first message's `value_type` (i.e. its topic) is not the same as one of the previous paragraphs' first messages' topics. This is done by comparing the prefixes of the `value_type`s.

The number of messages in a paragraph, as well as the number of paragraphs in the document, are also capped. In addition to these maximum numbers, the document planning process also includes an early-stopping condition: generation of either a paragraph or the document as a whole is stopped early if the newsworthiness values of all candidates fall below either an absolute or a relative threshold. These prevent the text from straying into minutiae.

The key feature of the work described in Paper IV are the limited assumptions made about the domain of the text. Generic works on document planning are few because the process is highly domain-specific. Most system descriptions either completely omit a description of the document planning component, describe very specific hand-engineered approaches, or use machine learning approaches. While the hand-engineered approaches are likely to overperform our method, that same costly hand-engineering also limits their usefulness to newsrooms (See Chapter 3).

Chapter 5

Evaluation

This chapter presents an evaluation of the computational methods described in Chapters 3 and 4. First, Section 5.1 discusses how news automation systems can and should be evaluated. The two following sections then present evaluations about the architecture and NLG approach as a whole, as well as of the document planning methods described in Section 4 in specific. The last section of this chapter then considers the general fitness for purpose of the methods described above.

5.1 How to evaluate news automation

In the following sections of this chapter, we describe evaluative efforts both contrasting computer-written news texts to human-produced texts, as well as evaluations where no comparison to human-produced text is made. Together, these evaluations cover both the view that the NLG outputs should be indistinguishable from the equivalent human product, as well as the – perhaps more business-valued – view that the important question is whether the produced texts are useful and valuable on their own merit. Notably missing from the following chapters are evaluations that employ automated metrics such as BLEU [73] or ROUGE [57], which are otherwise common in evaluation of natural language generation systems.

Automated metrics such BLEU and ROUGE are dependent on the existence of gold-standard texts against which the automatically generated texts are compared. These gold standard texts are expected to cover the whole space of "good" outputs. In the context of the studies evaluated in this chapter, no such corpus of gold-standard texts exists, and producing one is prohibitively expensive. Our experiences in conducting an evaluation based on comparisons to human-authored texts [65] indicate that producing

even a single such text by a domain expert takes significant time, invoking a very large cost. Furthermore, there is considerable variance in what factors different domain experts highlight from the data [65], indicating that the set of gold standard texts should be very large for any system input.

The natural language generation research community at large has also expressed concerns that evaluations based on – at least some – automated metrics are not necessarily reliable for scientific hypothesis testing [84]. For example, Novikova et al. [71] identified that "no metric produces an even moderate correlation with human ratings, independently of dataset, system, or aspect of human rating" while a survey by Celikyilmaz et al. [15] observes that "automatic metrics still fall short of replicating human decisions." Because "human assessment remains the most trusted form of evaluation in NLG" [50], the following sections evaluate the methods described above using human evaluations and qualitative methods.

5.2 Architecture and design

A successful journalistic NLG system must be (See Chapter 3, Paper I)

1. accurate;
2. transparent;
3. modifiable and transferable; and
4. sufficiently fluent.

Any operational journalistic NLG system also requires that sufficient amounts of topical data are readily available. This requirement, however, focuses on the operational setting of a specific system instance[1]. As such, the following analysis ignores it. This decision is supported by the systematic literature review of Grimme [40], which identifies data issues as "not [being] perceived as the most decisive" in the limited adoption of news automation in newsrooms.

The next subsections discuss the above technical requirements one at a time. In addition to Paper I, this section draws from our evaluative findings published separately [65], as well as a more complex human evaluation of another news automation system developed using the methods from Papers I, III and IV [56]. Reference is also made to several other related works [2, 81, 101] which include the author of this thesis as an author.

[1] For a broader discussion on viewing data as an operational concern, see [94].

5.2.1 Accuracy

The accuracy of the software is a fundamental requirement for news automation systems. By employing a rule-based approach to generation, we sidestep the hallucination problem, which is a major source of inaccuracy for neural generation approaches.

Further, the software implementation strives to keep the data immutable. This is done by separating the message objects into the message data derived from the underlying database (called a *fact* in Paper I) and the associated metadata. The critical data fields are then made immutable, preventing accidental modifications during generation. As a result, the system is highly robust against any unwanted modifications to the factual content of the story.

The immutability described above, however, still leaves the possibility of inaccuracies resulting from lexicalization effects. For example, an increase might be described as a decrease due to a templating mistake. Any such mistakes in the systems are, however, fundamentally addressable because the system is both transparent and modifiable, as discussed in the following subsections.

A more complex problem observed while evaluating the case study system developed in Paper I was caused by the simplistic approach to aggregation: in some cases, the aggregator (based on sentence prefixes) produced illogical and confusing output [65]. In later systems based on the same architecture, we addressed this issue by either employing a more complex aggregation method or by further limiting – even disabling – aggregation.

With the exception of the aggregation problems detailed above, the systems implementing the architecture have performed overall accurately. This is reflected in human evaluations of system implementations. Evaluating the case study system described in Paper I, non-expert human judges indicated that Finnish language news texts about the results of municipal elections were both credible ($\mu = 3.59$ on a 5-step Likert scale vs $\mu = 4.10$ for human-written articles) and representative of the underlying data ($\mu = 3.15$ on a 5-step Likert scale vs $\mu = 3.96$ for human-written articles) [65].

The architecture has since been re-evaluated in the context of a news automation system producing statistical news in multiple European languages [56]. On a 7-step Likert scale, expert evaluators viewed the software as producing news texts that were newsworthy ("Newsworthiness" median 5.5, mode 5) and not unuseful ("Usefulness" median 4.0, mode 4).

5.2.2 Transparency

The rule-based approach to news automation employed in the architecture provides a high level of inherent transparency: the templates and algorithms can be inspected, and their creators queried for insight. This stands in contrast to various "black box" approaches. Any decisions made by a software instantiation of the architecture described above can be traced back to a specific software component.

The transparency is further aided by the simplicity of the templating language used by the system. By employing a simple templating language, the templates can be inspected – and contributed to – by domain expert journalists in a much more direct fashion than with templating approaches that require significant linguistic knowledge.

Naturally, our approach presents a certain transparency-fluency tradeoff. On one hand, complex templating formats such as parse trees and long templates would likely produce, or at least allow for, more fluent outputs. On the other hand, simpler phrase-level templates such as those used by us, are easier to work with without linguistic knowledge. This potentially allows for a higher number and variety of said templates for any given time investment. Still, encoding any complex usage or agreement rules into the templates requires *some* linguistic knowledge irrespective of the templating language.

5.2.3 Modifiability and transferability

The modifiability of the proposed architecture is best demonstrated in practice through several published modifications made to systems embodying the architecture. We have demonstrated [2] how the pipeline-architecture accommodates new modules for automated generation of locator maps and graphed data with minimal effect on the rest of the pipeline. We have also demonstrated [81] how the architecture can accommodate neural processing modules for introducing increased variety into the generated language. In a third publication [3], we enhance the architecture with a system for producing more catchy and creative headlines. Finally, in both Paper IV and an unpublished manuscript [101] we demonstrate how the modular architecture allows for easy replacement of individual pipeline components with others without affecting the pipeline at large.

The transferability of the system to other domains is similarly best demonstrated in practice through several case studies. Taking the original system instantiation described in Papers I and III, which produces texts about elections, we have applied the same architecture – in fact reusing

much of the code base of the original case study system – to several new domains. These include national crime statistics [2], EU-wide economic data (Paper IV, as well as [81, 101]), and data obtained by applying natural language processing tools on historical newspaper corpora [77] and online news comments [104].

Together, these demonstrate that the architecture is both modifiable within domains, as well as extensible to other types of text domains beyond the election news domain of Papers I and III.

As with transparency, some of the design decisions that enhance transferability come at a cost of fluency. For example, the short templates defined in the simple templating language are easy to produce, allowing for quick bootstrapping in a new domain. On the other hand, as noted above, they come with limitations on e.g. linguistic information, thus limiting fluency.

5.2.4 Fluency

The evaluation of the case study system described in Paper I also asked non-experts about the fluency of the produced texts [65]. When presented with both randomly selected news texts written by the case study system, as well as equivalent texts produced by human expert journalists, the respondents showed a preference for the human-authored texts over the computer-generated texts. Measured on the "liking" facet (measuring "overall affective reaction" and exemplified by words such as "enjoyable", "interesting", "lively" and "pleasing") the respondents gave the human-authored texts a mean score of 3.98 on a 5-step Likert scale, in contrast to the mean score of 2.33 for computer-authored texts. The results for the "quality" facet (measuring the "degree or level of overall excellence of a news story" and exemplified by words such as "clear", "coherent", "comprehensive", "concise" and "well-written"), likewise measured on a 5-step Likert scale, mirror the above with human-written texts having a mean value of 3.96 compared to the computer-authored texts' mean of 2.58. On both facets, the differences were statistically significant.

At the same time, the respondents had at times trouble identifying whether the stories were computer- or human-produced [65]. When asked whether they believed the author of a text to be a human or a computer, the respondents incorrectly attributed a computer-generated story to a human in some 21% of evaluations. For the article where this happened the most, some 33% of the responses attributed the computer-generated text to a human. Misattributions were also made in the opposite direction: 10% of evaluations misattributed a human-authored story to a computer.

When the judges were given a choice to select what locale they read about (i.e. allowing them to self-personalize), they showed a statistically significant increase in the "liking" aspect for self-selected stories compared to the randomly selected texts [65]. The same was observed for the "quality" facet. While statically significant, the differences was not qualitatively very large at mean values 2.33 versus 2.68 for "liking" and 2.58 versus 2.75 for "quality". On the other hand, no statistically significant difference was observed in the perceived credibility between the preselected and self-selected texts.

In general, the free-text answers provided by the judges included phrases such as "boring", "listing-like", "abrupt", "robot-like", "stiff", "grammar mistakes" and "dry" as negative side, and "clear", "readable", "to the point", "objective" and "interesting" on the positive side [65]. Our interpretation of these results is that the computer-generated texts clearly fell behind human-authored texts in terms of their fluency, but at the same time, did reach the fluency threshold required for being acceptable.

The fluency aspect was also evaluated in another case study [56], where news professionals evaluated automatically produced statistical news texts. On a 7-point Likert-scale, the judges indicated that the texts were grammatical (median 6.0, mean 6), not unuseful (median 4.0, mean 4), and reusable (median 5.0, mode 5). The fluency of the texts was both varied and overall average (median 3.5, mode 5). In this latter evaluation, the texts were explicitly framed as being intended for newsroom-internal use, rather than as texts intended to be served directly to the readership.

5.3 Document planning

Evaluating individual components of a news automation pipeline is non-trivial when – as in the case of the work described in this thesis – no comprehensive corpora of gold standard outputs are available. As such, the methods described in Section 4 were evaluated as parts of complete case study systems.

5.3.1 Identifying newsworthy datapoints

In Paper III, we described a method for associating individual data points with a numeric estimate of "newsworthiness". This estimate was then combined with a relatively simple document structuring component in the election news case study system.

The evaluation of this case study system [65] focused on high-level attributes, of which perhaps the most relevant for content selection purposes

is the "representativeness" facet. This facet describes "a summary judgment of the extent to which the story is representative of the category of news. In other words, it is the answer to the following question: What is the probability that the story, taken as a whole, belongs to the class of entities that we call 'news'?" and is associated with lead words "important" and "relevant".

On a 5-step Likert scale, the non-expert evaluators (who represent the general news-reading public) evaluated the representativeness of automatically generated news texts at a mean value of 3.15. This is statistically significantly different ($p < 0.01$) from the mean of 3.96 they gave to texts written by expert human journalists. Unlike in some cases above, there was no statistically significant difference ($p = 0.55$) between the ratings for automatically generated texts selected for the judges, and those selected by the judges themselves.

An analysis of the free-text answers provided by the judges included both positive and negative aspects related to document planning. On the negative side, judges commonly mentioned "order" and "listing-like". On the positive side, they often mentioned words such as "facts", "clear", "most important", "objective" and "comprehensive". Overall, we identified "a trend of praising the fact-basedness and that the story is clear and to-the-point" while the negative aspects were more closely related to the language of the texts [65].

We interpret these results as indicating that the "newsworthiness" aspect of Paper III is fundamentally sound, but the content structuring aspect has room for improvement. This motivated the subsequent investigation of a more complex content structuring approach as described in Paper IV (see Section 4.2).

The above interpretation is supported by the evaluation of a news automation system producing statistical news in multiple European languages [56]. There, expert evaluators gave the produced texts a "newsworthiness" median score of 5.5 (mode 5) on a 7-step Likert scale.

5.3.2 Planning news reports

The work described in Paper IV was also evaluated using a human evaluation. Domain experts from the Finnish News Agency STT evaluated texts produced by both the proposed method and a simpler baseline method. The systems were identical except for the document planning components. As such, any differences in the evaluative results between these two systems can be attributed to the document planning components.

The judges indicated their agreement to five questions more directly tailored to document planning. The judges evaluated the proposed approach as statistically significantly superior in terms of having the contents of the text match the heading of the text (mean 4.40 vs 1.80; median 5 vs 2) and the document being coherent (mean 4.33 vs 1.60; median 5 vs 2), both measured on a 7-step Likert scale. The proposed method also appeared to outperform the simpler baseline in terms of the factors "the text lacks some pertinent information" and "the text contains unnecessary information", but these differences were not statistically significant once a correction for multiple comparisons was applied.

Queried whether the text length was suitable, the proposed method outperformed the simpler baseline (mean 2.93 vs 4.07; mean 3 vs 4) on a 5-step Likert scale ranging from 1 ("clearly too short") to 3 ("length is suitable") to 5 ("clearly too long").

Overall, we interpret these results to indicate that the proposed document planning method produces texts of at least sufficient coherence. As the approach is designed to minimize domain-specific knowledge, and is intended to be used as a baseline for quickly bootstrapping news automation systems in new domains, this result is very positive.

5.4 Fitness for purpose

Taken as a whole, the technological properties of the proposed methods match well with the key requirements of transparency, modifiability and transferability. However, the results in terms of text fluency and document planning require more careful analysis.

As identified by judges representing news consumers [65], the computer-authored texts appeared to be lacking in comparison to texts authored by expert journalists, indicating that they are not a drop-in replacement. At the same time, a sufficient level of fluency for baseline understandability was reached.

However, differences between the computer-generated and the human-authored texts need not be considered failures of the software development process. Instead, the nature of the automatically produced texts as exemplars of a new genre can be embraced by turning the difference into an advantage, as described in Paper II. As an example, consider an automatically produced news story that does not provide as much context and analysis as an equivalent human-written text would. Rather than framing this difference as a failure of the automation, the software's speed, accuracy and low cost can be emphasized, and the resulting text framed as a

new subgenre of the "news flash" where this lack of detailed analysis is the norm. The instantly available computer-authored text might not be as fluent as a detailed news-piece a domain expert would be able to draft in an hour, but in the meantime it is more fluent than *no text at all*.

This positive reframing can be taken further by incorporating the personalization aspects made possible by automation. Indeed, allowing the judges to self-personalize resulted in statistically significantly higher scores in some of the evaluated aspects.

News automation systems can also be useful in a context where they produce drafts of news texts targeted at journalists (rather than the general public) by the virtue of saving journalist time and highlighting to them potential news stories and angles. Indeed, news professionals viewed the texts produced by a system implementing the architecture described above as not unuseful (median 4.0, mean 4 on 7-step Likert scale) and believed that the texts could be reused as part of their work (median 5.0, mode 5 on 7-step Likert scale) [56]. We interpret these results as showing that the proposed methods produce sufficiently high quality texts for newsroom internal usage. This includes texts such as news alerts or drafts that function as a starting point for a human journalist.

A further potential usage case of automation combines these two: an automatically produced news flash, published instantly, ensures that the newsroom is "first on the scene", while also providing the human journalists with a foundation on which to build a more complex and analytical story. This question of how to best employ the methods described above is returned to in Chapter 7.

Chapter 6

Bias, authorship and ethics

The ongoing integration of the various automation and AI methods into newsrooms – under any label – is seen as a significant disruptor [19, 69]. As with any disruptive technology, increased automation is seen as having both positive and negative effects. In an analysis of interviews conducted with news editors [100], we previously identified beliefs that automation would provide efficiency of work, increase the output of the newsroom and allow for journalistic resources to be reallocated towards more investigate work. At the same time, the interviewees indicated several reservations regarding the state of the art of news automation technology, as well as the operational setting in which it was to be employed. Reservations about various ethical and societal concerns have also taken an increasing role in discussions regarding the use of automation and AI methods in society [19, 40]. In this chapter, we overview two important aspects of ethics of news automation: bias and authorship.

6.1 Bias and perceptions of bias

Traditional journalism is in choppy waters in terms of public trust. Despite being seen as incorporating an important objectivity norm [95] associated with terms such as "nonpartisanship", "facticity" and "balance" [66, p. 2], a 2017 survey found that only 28% of Americans believed that news media supported democracy well or very well [90]. A 2022 survey of Americans found similar results, with a mere 7% having "a great deal" and a further 27% having "a fair amount" of trust and confidence in media [8]. Surveying a broader range of nations, a 2022 report found that 42% of respondents trust "most news most of the time", with Finns exhibiting the highest numbers of trust at 69% [70].

In a survey of news media insiders [100], we identified that at least some media representatives believe automation has the potential to strengthen the news' trustworthiness in the eyes of the general public, and that the automated news texts represent "facts [..] and figures, not someone's manipulated interpretation" [100, p. 56]. These sentiments are mirrored by academics who see AI methods as a way of "rejuvenating public trust in journalism" [91].

But there is no reason to assume that a news automation system would be unbiased. As noted in Paper V, rule-based systems might incorporate various heuristics that, on closer inspection, exhibit (potentially very subtle) biases, while machine learning systems might learn to incorporate biases present in the human-produced training data.

On the content selection and document planning level, it is instructive to consider various gender biases shown by human journalists [29, 48, 96]. Mirroring these, an automated systems might prefer to discuss males over females in any of multiple contexts, for example showing a preference for male political candidates over female candidates. As an example of subtler bias, a news automation system might always mention that the suspect of the crime is an immigrant but leave the non-immigrant suspects unmarked.

Similarly, news automation systems might exhibit biases during the lexicalization stage, where they decide on the language – phrases, words – used to convey the information selected for inclusion in the text. For example, an increase in some statistic that is positive for the sitting government might be described as simply an "increase," while some other change that is negative for the sitting government might be described as "rocketing," even if the changes were largely equivalent. Alternatively, a news automation system might describe a 17-year-old perpetrator of a crime as either a "boy" or a "young man" depending on the background of the person.

Given that media professionals appear to generally acknowledge that humans indeed are (or at the least can be) biased, a question arises: why, then, do at least some assume that news automation systems would not be biased? In our view, this view is predicated on two assumptions.

The first assumption is that news automation removes the individual humans' effect from the news production process. This is an understandable assumption: automation and software provide facades of impartiality, masking the contribution of the individual human. However, in the case of rule-based systems, the rules are always produced by individuals and encode their beliefs and preconceptions, which might be biased in either blatant or very subtle ways. For machine learning models, the models are trained to mimic the actions of humans using training data obtained from

the actions of individual humans. This means that the models are also trained to mimic human biases. In both cases, the individual human remains involved and inseparable, even if hidden behind the facade of an impartial and uncaring computer.

The second assumption appears to be that removing the individual is sufficient to remove *all* human bias. This means, effectively, denying the effects of organizational and social factors, as well as larger social systems [82]. Even if we were able to somehow remove the effect of an individual from some machine learning model, that model would still remain fundamentally an artefact of our society and the organization that built it. Even a "neutral" football game report (i.e. one favouring neither team) will encode societal and group values such as what are the most important, newsworthy, elements in a game of football: do we celebrate those who scored the most goals, or those who supported their teammates by consistently building the game and passing the ball?

Even the concept of "bias" is valued. Any news production activity inherently involves questions of what information is conveyed to the reader [47] and how that information is "framed" [30]. The differentiation between the acceptable types of selectivity and framing, and those we consider harmful, is itself a question of human, organizational and societal values. If we assume that whatever values and framing devices are encoded into an automated systems created today are "fair" and "unbiased", we risk those frames and values becoming entrenched and axiomatic; something taken as granted and beyond criticism.

Acknowledging that news automation has the potential for bias is not just an important step in its ethical use. It also presents an opportunity. Researchers have developed methods for identifying biases in word embeddings [6, 38, 72], language models [97] and translation systems [17]. By acknowledging that news automation can be biased, it can instead be turned into a tool for researching bias in humans.

One might, for example, train a language model using human-authored news, and interrogate that model for biases. Alternatively, one might ask human journalists to help build a rule-based system for producing news texts about some domain, and then compare the news produced by the system to those produced by human journalists: any differences in focus or framing would then be of potential interest for identifying even subtle biases in the humans.

6.2 Authorship, responsibility and ownership

As news automation either diminishes or masks the influence and contribution of the individual human, difficult questions of authorship arise: who or what is the author of the text produced by a news automation system? This question can be approached from two distinct angles: from an instinctive angle of a general layperson, and from the point-of-view of the legal system which attributes authorship for purposes of responsibility and intellectual property rights.

Authorship

Investigating how the general public attributed authorship to a news text produced by the news automation system described in Paper I, Henrickson [45] found that of 500 respondents, 179 (35%) identified the system itself as the author of the text, 143 (29%) stated that it was not possible to assign authorship, and 90 (18%) assigned authorship to the team that developed the software. A total of 9 (2%) respondents assigned authorship to the party funding the creation of the news automation system. A further 72 respondents (14%) selected the option "other". Importantly, the responses indicate that relatively few people attribute the authorship to the humans who built the news automation system.

According to Henrickson [45], a common theme in the responses was an evocation of a parent-child metaphor, where the system was seen as a child and the developers as a parent, with the respondents then observing how human parents have no right to their children's creations. These results are mirrored in a later survey of the authorship debate by Henrickson [46]. They observe that discussions over authorship of machine-generated texts go back at least into the 1960s, with many scholars and authors explicitly denying the computer's authorship, while others distinguish between an "immediate" and an "ultimate" author.

Newsrooms, too, disagree on how to address the authorship of texts produced through news automation [67]. While some key figures attribute authorship of automatically produced texts to programmers, others consider the organization at large the author. However, while newsroom key figures *say* they reject the concept of the system itself as an author, at least some newsrooms attribute the algorithm on the byline.

Some professional guidance has begun to emerge in recent years. For example, the Finnish Council for Mass Media (CMM, the self-regulatory body of Finnish mass communication publishers and journalists) published a statement on news automation and personalization in 2019 [33].

In the statement, the CMM does not explicitly take a stance on authorship [33]. However, they clarify their previous guidance, explaining that ethical guidelines for journalists should be viewed as applying to "digital service developers". This acknowledges the influence and importance of the developers that created the news automation system, and indicates that at least some of the authorship should be viewed as belonging to them.

Second, they recommend that any texts including "to an essential extent" content produced by news automation should be clearly marked as such [33]. This indicates that in the CMM's view, even in a context where a human journalist collaborates with a news automation system, some of the authorship lies outside of said human journalist. In other words, news automation is seen as standing aside from other newsroom technologies: one would not be expected to explicitly mark down that a piece of news text was spell-checked using a computer system.

Responsibility

Fundamentally associated with authorship is the author's responsibility for the text they produce. In a journalistic context, it might be tempting to short-circuit this discussion by stating that every newsroom has an editor-in-chief who holds final accountability for the output of the newsroom, and thus no further consideration is needed. In this regard, the use of news automation is in a way no different from the editor hiring a new journalist to work under their supervision. But even if the editor-in-chief holds the *final* responsibility, that does not absolve a human journalist working under them from all responsibility. Why, then, should the analysis stop at the editor-in-chief in the case of a news automation system?

Furthermore, placing all the responsibility over highly technical computer systems on a single individual – who is unlikely to completely understand the details of complex computer systems – is ethically problematic in itself. Intuitively, one would expect the responsibility to also somehow reside "closer" to the system. Yet the proximate group of individuals (the programmers and journalists who built the system) for any news automation system can often be so large that any responsibility would be diluted: no individual is truly responsible, except for the far-removed person who is responsible for everything.

CMM's 2019 guidance [33] acknowledges the complexity of the situation. It states that all decisions regarding journalistic content, including those made by news automation, must be retained wholly within the editorial office, and that the ultimate responsibility stays within the editorial staff and the editor-in-chief. This statement is necessary only if, in their view,

the use of automation had the potential to move some of the control outside of the editorial office, indicating that at least some authorship lies within the developers.

As for solving these ethical problems in practice, CMM's statement is not very helpful. CMM only states that "media outlets should have sufficient understanding of the effect of algorithmic tools on content" [33]. How this "sufficient understanding" would be obtained is less clear, as is what is "sufficient".

Copyright and intellectual property rights

Authorship is also intimately tied to ownership through copyright and associated intellectual property rights. Overviewing the legal field, Bridy [9] observed that the fundamental problem faced is that "[t]he author of a procedurally generated artwork is, for all intents and purposes, *another copyrighted work*". They conclude that, because the proximate author (the computer program) has no legal personhood, the copyright should then intuitively (even if the relevant statutes might not be clear) transfer to "the author of the author", i.e. the person who authored the computer program.

In the context of EU law, legal scholars have identified "clear indications in legislative material as well as case law suggest[ing] that the concept of work and authorship are dependent on human efforts" [92, p. 173] and that "according to historically undisputed anthropocentric copyright doctrine [..] only works created by natural persons enjoy copyright protection" [76, p. 217]. In terms of related rights, patent law is even more clear, explicitly stating that "the inventor designated in a European patent *must be a natural person*" [4, p. 200] (emphasis in original).

Assuming that the credit and profit go hand-in-hand with the blame and responsibility, the laws appear to support an analysis where the authorship lies in some collection of humans most instrumental to the creation of the news automation system. In analysing how the European copyright system incentivizes certain types of actions relating to news automation [76], we identified that this status-quo appears to align well with at least rule-based systems: both the credit and the responsibility can be assigned to the authors of the rules.

More complex methods of automation, however, begin to strain the current legal framework. How much authorship can be attributed to a person who trains a large language model on data scraped from the internet? How should we view the act of creating text with a language model trained solely with the outputs of a single journalist? For news text in specific, the Press Publishers' Right of the EU Digital Single Market Directive does not re-

quire human authorship for a short-term protection, thus short-circuiting the discussion in part [76]. However, the protection afforded by the Press Publishers' Right is distinctly shorter than copyright and does not extend beyond news, leaving these questions partially unanswered.

All-told, the authorship question – and the associated questions of credit and blame – remains complex and without final answers. They are also unlikely to be resolved for good while the larger debate on ethics of AI continues.

Chapter 7

Discussion

In this chapter, we first discuss how news automation should and could be employed in the light of both the evaluative results obtained in Chapter 5 and the ethical considerations discussed in Chapter 6. This is followed by a discussion of potential limitations of this thesis.

7.1 How and where to employ news automation

News automation methods have already established themselves well in certain domains where plentiful input data is available and the texts being produced are either highly standardized (e.g. weather and earnings reporting) or can be distilled to highly archetypal stories (e.g. some sports and elections). For these domains, the volume of produced news is sufficiently high to allow for the creation of highly customised (and thus costly) systems that produce fluent and high-quality news texts to be served directly to the customers.

As automation brings the amortized cost of a single news story lower, stories with smaller potential readerships become financially feasible. A story about some minor league football game might not make financial sense when written by a human journalist, but can be a sensible target for automation. Indeed, automated football news have been investigated by several newsrooms especially in the Nordic countries [59]. These types of niche news allow the targeting of the "long tail" of news [65].

In the above cases news automation systems are often understood to be complete replacements for the human journalist. Thus, the assumption is that the texts are of near-human fluency. However, as discussed in Chapter 5 and Paper II, there are other ways of integrating news automation into the newsrooms, allowing for the use of medium-fluency textual content.

First, news automation methods can be framed as producing texts of a genre that de-emphasizes fluency and instead values other attributes associated with automation, such as speed, predictability, price and transparency. This can be done by, for example, framing the news texts as "*breaking news alerts*". The methods described in this thesis are a good fit for such a task. They make it possible to monitor a large amount of data sources while minimizing the cost of tailoring required by each data source. This, in turn, allows the "net of automation" to capture increasingly rare or surprising events.

Second, news automation can be used internally to the newsroom. Rather than having news automation methods target the general audiences, the automatically produced texts can be targeted at journalists. In this scenario, news automation produces "news alerts" that highlight potential news stories for the expert human journalists, while concurrently serving as a first draft of the story. In doing so, they would presumably bring down the response time from an event to publication. If the time to publication is crucial, human journalists could even publish the computer-authored "first draft" as a breaking news alert, buying the human more time while still being "first on the scene".

Third, news automation can be framed as a co-author. As a co-author, news automation can produce shorter textual components that can be integrated into more complex human-authored stories. In these types of co-authored texts, the role of news automation could be to tailor and personalize the text, for example by inserting segments that closely relate to the reader. For example, a human-authored news article about the latest national unemployment figures could contain an automatically produced section about the reader's area in specific. In the best case, such collaboration allows for a "best of both worlds" situation, where the resulting texts are more personalized than a human journalist would be able to produce alone, and more fluent and human than a news automation system would be able to produce alone. The methods presented in this thesis should lower the cost of producing these types of systems.

Returning to the landscape of automation in newsrooms (Section 2.2), these approaches to integrating news automation focus on different related technologies. By employing news automation as a producer of news alerts or drafts internal to the newsroom, the similarities with Computer-Assisted Reporting are emphasized: a well-established news automation system of this type would likely be eventually viewed the same way as word processors and search engines are viewed now. On the other hand, when news automation is used to produce consumer-facing texts, the Computational

Journalism and Data Journalism aspects are emphasized: those who produce news automation systems of this type could be seen as a new type of journalist, producing not text but systems that produce text.

Newsroom-internal and collaborative uses of news automation (i.e. producing texts that combine human-authored and computer-authored elements) also go towards solving some of the ethical issues associated with the use of automation. If the human journalist has significant say in how the news automation system functions, for example by contributing templates and deciding what data to include, they can retain control and an understanding of the automation. This allows them to claim both responsibility and credit for the contributions of the automated system in an ethically sound manner. In terms of responsibility, this would go a long way towards matching the guidance and ethical guidelines of industry self-regulatory bodies such as the Finnish Council for Mass Media.

Collaborative and newsroom-internal news automation would also sidestep – at least in part – some of the issues regarding copyright and associated intellectual property rights. Even if the automatically produced part of the text was later found to not be under copyright, the human collaborator's contribution would presumably be under copyright to the degree that most news texts are. As noted above, rule-based methods such as those described in this thesis are also much easier to align with existing intellectual property rights case-law and legislation than, for example, methods based on large language models or other neural text generation methods.

Framing news automation as a tool rather than as a replacement might also provide other benefits. Most notably, it might go some way towards softening some of the concerns regarding the effects of automation on the newsroom work force.

7.2 Limitations

As with any academic study, this study has several potential limitations. Discussing each research objective at a time, we have identified the following potential limitations:

RO1: Identify (a) requirements and (b) a high-level architecture for news automation A key assumption in the qualitative analysis of the proposed architecture is that the requirement analysis stage correctly identified the necessary requirements. If one or more requirements were missed – despite our best efforts – it is possible that the proposed architecture would not meet all the real-world requirements imposed by the

journalistic context. Furthermore, as the architecture's alignment with the requirements was evaluated qualitatively, it is possible that some important consideration was overlooked.

There are also possible concerns of external validity, primarily on the degree to which the results generalize outside of these studies. It is possible that the architecture makes some implicit assumptions we have not yet recognized, and which were not encountered in the various case studies. If so, these assumptions might limit the architecture's generalizability.

In terms of the multiple human evaluations, the use of human evaluations is susceptible to various effects that reduce the validity of the evaluations. For example, some of the evaluations cited had a limited number of participants. It is also possible that those who are inclined to participate in such studies would be either interested in and thus susceptible to respond more positively to news automation, or alternatively concerned by it, and thus susceptible to respond negatively to it. In the only study where participants were asked whether they were familiar with news automation [65], we identified that the average ratings differed by 0.13 to 0.25 between different groups of respondents based on amount of news consumption and familiarity with news automation. Finally, it is possible that some of the evaluators would have either misunderstood the questions, or intentionally answered incorrectly or without thinking for any of myriad reasons.

RO2: Identify broadly applicable methods for planning and structuring documents for news automation The content selection and document planning methods were evaluated as parts of complete news automation systems. Because of this, it is possible that some properties of the broader system would have affected how the evaluators perceived the texts they were evaluating. The above concerns regarding human evaluations also apply to this research objective. Finally, it is possible that the assumption about the existence of a hierarchical labeling scheme limits the generalizability of the proposed method.

RO3: Identify (a) how and why news automation systems can be biased, and (b) what could explain a hesitancy to accept news automation's potential for bias Concerns remain regarding the practical application of the proposed use of news automation to detect biases in human journalists. It remains possible that it would not yield practical results in some contexts, or alternatively would fail to catch certain types of biases. Further study on the use of news automation to diagnose biases in humans is required to establish the limits of the proposed method.

Chapter 8

Conclusions

In this chapter, we first review the research objectives identified at the start of this introductory part of the thesis, as well as the main results relating to them. Finally, we describe some potential avenues for future work.

8.1 Revisiting the research objectives

In this thesis we have sought to answer a singular research question: *how to best conduct data-to-text natural language generation for factuality-emphasizing domains, such as journalistic reporting.* To help answer this broad question, we identified three main research objectives (ROs):

RO1: Identify (a) requirements and (b) a high-level architecture for news automation

RO2: Identify broadly applicable methods for planning and structuring documents for news automation

RO3: Identify (a) how and why news automation systems can be biased, and (b) what could explain a hesitancy to accept news automation's potential for bias

RO1: Identify (a) requirements and (b) a high-level architecture for news automation The first research objective is answered through a combination of Papers I and II (Chapter 3). We identified key domain and genre conventions relating to journalism and then rephrased those conventions as system and architecture requirements for accuracy, fluency, system transparency, modifiability and transferability. In addition, we identified the availability and topicality of data as key in determining where news automation can be applied.

Based on these requirements, we adapted the modular NLG architectures described by others for news automation. The proposed architecture prioritizes modularity, transparency and transferability by separating much of the domain-dependent processing into general software components parametrized by domain-specific information.

As evidenced by the several case studies and evaluations (Chapter 5), the architecture fulfills the requirements sufficiently well to be useful to newsrooms by producing texts for journalists, as well as alone in certain contexts where properties such as speed, predictability, price and transparency are highly valued.

RO2: Identify broadly applicable methods for planning and structuring documents for news automation The second research objective is answered through a combination of Papers III and IV (Chapter 4). We operationalized the concept of "newsworthiness" and provided a method for assigning individual data points in large datasets a numeric approximation of their newsworthiness. We also gave a domain-independent method for producing document plans through a combination of the above numeric estimates of newsworthiness and hierarchical metadata associated with each data point.

Our evaluations (Chapter 5, see also discussion in Chapter 7) indicate that the methods' performance is sufficiently high to be useful as long as the produced texts are framed suitably. The methods successfully limit the required domain knowledge, thus fulfilling their design goal and answer a key weakness in previous works.

RO3: Identify (a) how and why news automation systems can be biased, and (b) what could explain a hesitancy to accept news automation's potential for bias The third and final research objective is answered through Paper V (Chapter 6). We identified that news automation can be biased, and that the bias can manifest through both content selection and document planning, as well as the language of the text. We further described how these biases can be ultimately traced to the humans producing the system or its training data.

We then contrasted these findings to beliefs that news automation would be somehow inherently unbiased, identifying two assumptions that might lead one to such a conclusion. The first of these is an assumption that the use of automation removes the influence of an individual human. The second is that this removal of the individual removes all human bias. Finally, we proposed to use news automation systems to identify biases in human

journalists by producing news automation systems and interrogating where, how and why their outputs differ from those of human journalists.

Answering the research question Based on the answers and insights obtained through these three research objectives, we conclude that the rule-based approaches to news automation described in this thesis can be used for news automation purposes successfully. Since the methods are designed from ground-up to be as domain-independent as possible, these results indicate that they are also suited for report generation in other similar domains. This is further demonstrated by our applications of the proposed methods to the production of reports from historical newspaper archives [77] and newspaper comments [104]. In total, the methods described in this thesis go towards bridging the gap between previous natural language generation methods, and requirements for transferability and modifiability while maintaining accuracy.

8.2 Future work

In terms of building on the work described in this thesis, we see two especially promising research avenues. The first relates to improving on the methods described in this thesis, while the latter is about extending their use to new domains.

The first interesting research avenue is to consider the various ways in which machine learning methods could be combined with the methods discussed in this thesis. Machine learning methods present an enticingly high "quality ceiling" especially in terms of output fluency. Because of this, it would be beneficial if rule-based methods could be integrated with machine learning to produce systems that have the "quality floor" of rule-based systems and the quality ceiling of machine learning system. Our early works in introducing lexical variety [81] demonstrate one way in which machine learning components might be introduced into rule-based systems.

In selecting how to integrate machine learning approaches to otherwise rule-based systems, care must be taken to ensure that the failure modes of the machine learning systems are acceptable in the context wherein the texts are used. For example, a neural module that rephrases the output of an otherwise rule-based system would likely improve the fluency and variety of the output, but concurrently means accepting the possibility of errors similar to those observed from end-to-end neural NLG systems.

On the other hand, it appears likely that neural methods could be integrated into the document planning stage relatively safely. For example,

a neural module could be trained to select or order messages. As such a module would operate on outputs that are known to be good (i.e. messages obtained from the underlying data), and does not actually *modify* the data, it would be unable to produce outputs that are unfaithful to the input data. Thus, the worst-case scenario is simply that the produced story is "uninteresting" or difficult to follow, rather than containing outright factual inaccuracies. We are currently finalizing a publication on an investigation of one such neural document planning method [101].

Machine learning methods could also be useful in *building* NLG systems as long as humans have the ability to check and refine their contributions. For example, machine translation models could be used to produce (rough) translations of template files, making it easier to translate systems to other languages. Similarly, rephrasing models could be used to produce alternative versions of human-authored templates. Recent works by others have also investigated how to extract templates from large language models [107]. In such scenarios, human verification of the models' outputs would allow for any invalid contributions to be rejected or corrected by hand.

As a second interesting research avenue, the methods described herein could be applied to other domains, as we have already done with historical newspaper archives [77] and online news comments [104]. Personally, the author of this thesis finds our pilot of applying very simple report generation methods to learning analytics [55] promising: surely there is much that we could tell students about how to improve their studying habits. In any case, a field of plenty remains for future work.

References

[1] Asif Agha. Register. *Journal of Linguistic Anthropology*, 9(1/2):216–219, 1999.

[2] Rola Alhalaseh, Myriam Munezero, Miika Leinonen, Leo Leppänen, Jari Avikainen, and Hannu Toivonen. Towards data-driven generation of visualizations for automatically generated news articles. In *Proceedings of the 22nd International Academic Mindtrek Conference*, pages 100–109. Association for Computing Machinery, 2018.

[3] Khalid Alnajjar, Leo Leppänen, and Hannu Toivonen. No time like the present: Methods for generating colourful and factual multilingual news headlines. In *Proceedings of the 10th International Conference on Computational Creativity*, pages 258–265. Association for Computational Creativity, 2019.

[4] Rosa Maria Ballardini and Robert van den Hoven van Genderen. Artificial intelligence and intellectual property rights: The quest or plea for artificial intelligence as a legal subject. In *Artificial Intelligence and the Media*, chapter 8, pages 192–214. Edward Elgar Publishing, 2022.

[5] Kevin G Barnhurst and John Nerone. Journalism history. In *The Handbook of Journalism Studies*, pages 37–48. Routledge, 2009.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357. Curran Associates, Inc., 2016.

[7] Laurent Bourbeau, Denis Carcagno, Eli Goldberg, Richard Kittredge, and Alain Polguere. Bilingual generation of weather forecasts in an

operations environment. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*, pages 318–320, 1990.

[8] Megan Brenan. Americans' trust in media remains near record low. *Gallup*, 2022. https://news.gallup.com/poll/403166/americans-trust-media-remains-near-record-low.aspx. Accessed 2022-12-19.

[9] Annemarie Bridy. Coding creativity: Copyright and the artificially intelligent author. *Stanford Technology Law Review*, pages 1–28, 2012.

[10] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. From niches to riches: Anatomy of the long tail. *MIT Sloan Management Review*, 47(4):67–71, 2006.

[11] Erik Brynjolfsson, Yu Jeffrey Hu, and Michael D Smith. The longer tail: The changing shape of Amazon's sales distribution curve. *Social Science Research Network (SSRN)*, 2010. SSRN Preprint.

[12] Matt Carlson. The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, 3(3):416–431, 2015.

[13] Matt Carlson. The information politics of journalism in a post-truth age. *Journalism Studies*, 19(13):1879–1888, 2018.

[14] Donald O Case and Lisa M Given. *Looking for information: A survey of research on information seeking, needs, and behavior*. Emerald Group Publishing Limited, 4 edition, 2016.

[15] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[16] Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109. Association for Computing Machinery, 2019.

[17] Chloe Ciora, Nur Iren, and Malihe Alikhani. Examining covert gender bias: A case study in Turkish and English machine translation models. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 55–63. Association for Computational Linguistics, August 2021.

[18] Mark Coddington. Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital Journalism*, 3(3):331–348, 2015.

[19] Hannes Cools, Baldwin Van Gorp, and Michael Opgenhaffen. Where exactly between utopia and dystopia? A framing analysis of AI and automation in US newspapers. *Journalism*, OnlineFirst, 2022.

[20] Robert Dale. Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26(4):481–487, 2020.

[21] Laurence Danlos. Écriture automatique. *Les Cahiers de l'INRIA-La Recherche*, 443 Juillet-Août 2010, 2010.

[22] Kees Van Deemter, Mariët Theune, and Emiel Krahmer. Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31(1):15–24, 2005.

[23] Mark Deuze. What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism*, 6(4):442–464, 2005.

[24] Nicholas Diakopoulos. Computational news discovery: Towards design considerations for editorial orientation algorithms in journalism. *Digital Journalism*, 8(7):945–967, 2020.

[25] Laurence Dierickx. The social construction of news automation and the user experience. *Brazilian Journalism Research*, 16(3):432–457, 2020.

[26] Konstantin Nicholas Dörr. Mapping the field of algorithmic journalism. *Digital Journalism*, 4(6):700–722, 2016.

[27] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. Findings of the E2E NLG Challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328. Association for Computational Linguistics, 2018.

[28] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, 59:123–156, 2020.

[29] Susan Tyler Eastman and Andrew C Billings. Sportscasting and sports reporting: The power of gender bias. *Journal of Sport and Social Issues*, 24(2):192–213, 2000.

[30] Robert M Entman. Framing: Towards clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

[31] Alexander Fanta. Putting Europe's robots on the map: Automated journalism in news agencies. Fellowship paper, Reuters Institute for Study of Journalism, 2017.

[32] Thiago Castro Ferreira, Chris van der Lee, Emiel Van Miltenburg, and Emiel Krahmer. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, 2019.

[33] Council for Mass Media. Statement on marking news automation and personalization. https://www.jsn.fi/en/lausumat/statement-on-marking-news-automation-and-personalization/, 2019. Accessed 2022-09-08.

[34] Johan Galtung and Mari Holmboe Ruge. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1):64–91, 1965.

[35] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[36] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo,

Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120. Association for Computational Linguistics, August 2021.

[37] Eli Goldberg, Norbert Driedger, and Richard I Kittredge. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53, 1994.

[38] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614. Association for Computational Linguistics, June 2019.

[39] Andreas Graefe. Guide to automated journalism. Report, Tow Center for Digital Journalism, Columbia University, 2016.

[40] Meike Grimme. Factors influencing the rejection of automated journalism: A systematic literature review. *Nordic Journal of Media Management*, 2(1):3–21, 2021.

[41] Astrid Gynnild. Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets. *Journalism*, 15(6):713–730, 2014.

[42] Lauri Haapanen. Media councils and self-regulation in the emerging era of news automation. Report, Julkisen sanan neuvosto, 2020.

[43] James T Hamilton and Fred Turner. Accountability through algorithm: Developing the field of computational journalism. In *Report from the Center for Advanced Study in the Behavioral Sciences, Summer Workshop*, pages 27–41, 2009.

[44] Tony Harcup and Deirdre O'neill. What is news? News values revisited (again). *Journalism Studies*, 18(12):1470–1488, 2017.

[45] Leah Henrickson. Natural language generation: Negotiating text production in our digital humanity. In *Proceedings of the Digital Humanities Congress 2018*, 2019.

[46] Leah Henrickson. Authorship in computer-generated texts. In *Oxford Research Encyclopedia of Literature*. Oxford University Press, 2020.

[47] C Richard Hofstetter and Terry F Buss. Bias in television news coverage of political events: A methodological analysis. *Journal of Broadcasting & Electronic Media*, 22(4):517–530, 1978.

[48] Marc Hooghe, Laura Jacobs, and Ellen Claes. Enduring gender bias in reporting on political elite positions: Media coverage of female MPs in Belgian news broadcasts (2003–2011). *The International Journal of Press/Politics*, 20(4):395–414, 2015.

[49] Robert Howard. UTOPIA: Where workers craft new technology (1985). In *Perspectives on the computer revolution*, pages 341–349. Ablex Publishing Corporation, 2 edition, 1989.

[50] David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182. Association for Computational Linguistics, 2020.

[51] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 44(12):1–38, November 2022.

[52] Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski, and Filip Ginter. Template-free data-to-text generation of Finnish sports news. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 242–252. Linköping University Electronic Press, September–October 2019.

[53] Karen Kukich. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference*, pages 145–150. Association for Computational Linguistics, 1983.

[54] Karen Kukich. Knowledge-based report generation: A technique for automatically generating natural language reports from databases. *ACM SIGIR Forum*, 17(4):246–250, 1983.

[55] Leo Leppänen, Arto Hellas, and Juho Leinonen. Piloting natural language generation for personalized progress feedback. In *2022 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2022.

[56] Leo Leppänen, Hannu Toivonen, Eliel Soisalon-Soininen, and Matej Martinc. Final evaluation report on multilingual text generation technology. Deliverable D5.7 of the EMBEDDIA research project, 2022. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e9080ce3&appId=PPGMS.

[57] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004.

[58] Carl-Gustav Lindén. Decades of automation in the newsroom: Why are there still so many jobs in journalism? *Digital journalism*, 5(2):123–140, 2017.

[59] Carl-Gustav Lindén. Algorithms are a reporter's new best friend: News automation and the case for augmented journalism. In *The Routledge Handbook of Developments in Digital Journalism Studies*, pages 237–250. Routledge, 2018.

[60] Carl-Gustav Lindén, Hanna Tuulonen, Asta Bäck, Nicholas Diakopoulos, Mark Granroth-Wilding, Lauri Haapanen, Leo Leppänen, Magnus Melin, Tom Moring, Myriam Munezero, Stefanie Sirén-Heikel, Caj Södergård, and Hannu Toivonen. News automation: The rewards, risks and realities of 'machine journalism'. Report, WAN-IFRA, 2019.

[61] Per Linell. *Approaching dialogue: Talk, interaction and contexts in dialogical perspectives*, volume 3 of *IMPACT: Studies in Language, Culture and Society*. John Benjamins Publishing, 1998.

[62] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4881–4888. AAAI Press, 2018.

[63] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[64] Brian McNair. After objectivity? Schudson's sociology of journalism in the era of post-factuality. *Journalism Studies*, 18(10):1318–1333, 2017.

[65] Magnus Melin, Asta Bäck, Caj Södergård, Myriam D Munezero, Leo J Leppänen, and Hannu Toivonen. No landslide for the human journalist - An empirical study of computer-generated election news in Finland. *IEEE Access*, 6:43356–43367, 2018.

[66] David TZ Mindich. *Just the facts: How "objectivity" came to define American journalism.* NYU Press, 2000.

[67] Tal Montal and Zvi Reich. I, robot. You, journalist. Who is the author? Authorship, bylines and full disclosure in automated journalism. *Digital Journalism*, 5(7):829–849, 2017.

[68] Amit Moryossef, Yoav Goldberg, and Ido Dagan. Step-by-Step: Separating planning from realization in neural data-to-text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2267–2277. Association for Computational Linguistics, 2019.

[69] Nic Newman. Journalism, media and technology trends and predictions 2018. Report, Reuters Institute for the Study of Journalism, 2018.

[70] Nic Newman, Richard Fletcher, Craig T. Robertson, Kirsten Eddy, and Rasmus Kleis Nielsen. Reuters Institute Digital News Report 2022. Report, Reuters Institute for the Study of Journalism, 2022.

[71] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252. Association for Computational Linguistics, September 2017.

[72] Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 446–457. Association for Computing Machinery, 2020.

[73] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In

Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[74] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186. Association for Computational Linguistics, 2020.

[75] David L Parnas. On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12):1053–1058, 1972.

[76] Taina Pihlajarinne, Alexander Thesleff, Leo Leppänen, and Sini Valmari. The European copyright system as a suitable incentive for AI-based journalism? In *Artificial Intelligence and the Media*, chapter 9, pages 215–239. Edward Elgar Publishing, 2022.

[77] Lidia Pivovarova, Axel Jean-Caurant, Jari Avikainen, Khalid Alnajjar, Mark Granroth-Wilding, Leo Leppänen, Elaine Zosa, and Hannu Toivonen. Personal research assistant for online exploration of historical news. In *European Conference on Information Retrieval*, pages 481–485. Springer, 2020.

[78] Horst Pöttker. News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

[79] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6908–6915. AAAI Press, 2019.

[80] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035. Association for Computational Linguistics, 2019.

[81] Miia Rämö and Leo Leppänen. Using contextual and cross-lingual word embeddings to improve variety in template-based NLG for automated journalism. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 62–70. Association for Computational Linguistics, 2021.

[82] Stephen D Reese and Pamela J Shoemaker. A media sociology for the networked public sphere: The hierarchy of influences model. *Mass Communication and Society*, 19(4):389–410, 2016.

[83] Ehud Reiter. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104, 2007.

[84] Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018.

[85] Ehud Reiter. Challenges are same for neural and rule NLG. https://ehudreiter.com/2021/11/08/challenges-same-neural-rule-nlg/, 2021. Accessed 19-12-2022.

[86] Ehud Reiter. NLG systems must be customisable. https://ehudreiter.com/2021/02/17/nlg-systems-must-be-customisable/, 2021. Accessed 19-12-2022.

[87] Ehud Reiter. What are the problems with rule-based NLG? https://ehudreiter.com/2022/01/26/problems-with-rule-based-nlg/, 2022. Accessed 19-12-2022.

[88] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[89] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.

[90] Zacc Ritter and Jeffrey M Jones. Media seen as key to democracy but not supporting it well. *Gallup*, 2018. Available online at https://news.gallup.com/poll/225470/media-seen-key-democracy-not-supporting.aspx. Accessed 2022-12-19.

[91] Christopher Robertson and Anthony Ridge-Newman. The potential of artificial intelligence to rejuvenate public trust in journalism. In *Futures of Journalism*, chapter 9, pages 127–142. Springer, 2022.

[92] Ole-Andreas Rognstad. Creations caused by humans (or robots)? artificial intelligence and causation requirements for copyright protection in EU law. In *Artificial Intelligence and the Media*, chapter 7, pages 172–191. Edward Elgar Publishing, 2022.

[93] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[94] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15. Association for Computing Machinery, 2021.

[95] Michael Schudson. The objectivity norm in American journalism. *Journalism*, 2(2):149–170, 2001.

[96] Joseph Schwartz. Whose voices are heard? Gender, sexual orientation, and newspaper sources. *Sex Roles*, 64(3):265–275, 2011.

[97] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293. Association for Computational Linguistics, August 2021.

[98] Pamela J Shoemaker and Stephen D Reese. *Mediating the message*. White Plains, NY: Longman, 1996.

[99] Herbert A. Simon. What computers mean for man and society (1977). In *Perspectives on the computer revolution*, pages 445–458. Ablex Publishing Corporation, 2 edition, 1989.

[100] Stefanie Sirén-Heikel, Leo Leppänen, Carl-Gustav Lindén, and Asta Bäck. Unboxing news automation: Exploring imagined affordances of automation in news journalism. *Nordic Journal of Media Studies*, 1(1):47–66, 2019.

[101] Eliel Soisalon-Soininen and Leo Leppänen. Neural document planning for statistical news alerts without aligned training data. Unpublished manuscript.

[102] Peri Tarr, Harold Ossher, William Harrison, and Stanley M Sutton Jr. N degrees of separation: Multi-dimensional separation of concerns. In *Proceedings of the 21st International Conference on Software engineering*, pages 107–119. Association for Computing Machinery, 1999.

[103] Elizabeth A Thomson, Peter RR White, and Philip Kitley. "Objectivity" and "hard news" reporting across cultures: Comparing the news report in English, French, Japanese and Indonesian journalism. *Journalism Studies*, 9(2):212–228, 2008.

[104] Hannu Toivonen, Leo Leppänen, Aleš Žagar, and Marko Robnik-Šikonja. Final evaluation report on multilingual text generation technology. Deliverable D3.5 of the EMBEDDIA research project, 2022. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5df281c51&appId=PPGMS.

[105] Arjen Van Dalen. The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice*, 6(5-6):648–658, 2012.

[106] Peter White. Death, disruption and the moral order: The narrative impulse in mass-media 'hard news' reporting. In *Genres and Institutions: Social Processes in the Workplace and School*, chapter 4, pages 101–133. Cassell, 1997.

[107] Tianyi Zhang, Mina Lee, Lisa Li, Ende Shen, and Tatsunori B Hashimoto. TempLM: Distilling language models into template-based generators. *arXiv preprint arXiv:2205.11055*, 2022.