Finding and Expressing News From Structured Data

Full Paper

Leo Leppänen
University of Helsinki
Department of Computer Science
Helsinki, Finland
leo.leppanen@helsinki.fi

Myriam Munezero
University of Helsinki
Department of Computer Science
Helsinki, Finland
myriam.munezero@helsinki.fi

Stefanie Sirén-Heikel
University of Helsinki
Swedish School of Social Sciences
Helsinki, Finland
stefanie.siren-heikel@helsinki.fi

Mark Granroth-Wilding
University of Helsinki
Department of Computer Science
Helsinki, Finland
mark.granroth-wilding@helsinki.fi

Hannu Toivonen
University of Helsinki
Department of Computer Science
Helsinki, Finland
hannu.toivonen@helsinki.fi

ABSTRACT

In the age of increasing floods of information, finding the news signals from the noise has become increasingly resource and time intensive for journalists. Generally, news media companies have the important role of filtering and explaining this flood of information to the public. However, with the increase in availability of data sources, human journalists are unable to catch and report on all the news. This limitation, coupled with the need for media companies to continuously provide value to news readers, calls for automated solutions, such as automatically generating news from data. In order to support the journalists and media companies, and to provide value to audiences, this work proposes approaches for automatically finding news or newsworthy events from structured data using statistical analysis. Utilizing a real natural language news generation system as a case study, we demonstrate the feasibility and benefits of automating those processes. In particular, the paper reveals that through automation of the news generation process, including the generation of textual news articles, a large amount of news can be expressed in digestible formats to audiences, at varying local levels, and in multiple languages. In addition, automation allows the audience to tailor or personalize the news they want to read. Results of this work thus support and broaden the news offering and experiences for both media companies and the public.

CCS CONCEPTS

• **Applied computing** → **Media arts**; • **Information systems** → *Information systems applications*; *Content analysis and feature selection*; • **Social and professional topics** → *Automation*;

KEYWORDS

automated journalism, newsworthiness determination, news values, natural language generation

ACM Reference format:
Leo Leppänen, Myriam Munezero, Stefanie Sirén-Heikel, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Finding and Expressing News From Structured Data. In *Proceedings of AcademicMindtrek'17, Tampere, Finland, September 20–21, 2017,* 10 pages.
DOI: 10.1145/3131085.3131112

AcademicMindtrek'17, Tampere, Finland

978-1-4503-5426-4/17/09…$15.00
DOI: 10.1145/3131085.3131112

1 INTRODUCTION

News gives us meaningful knowledge of events, functioning as a valuable asset to be shared with others. Since the advent of modern news media, the creation of news has followed a similar set of characteristics around the globe, based on a occupational ideology defining the profession [11]. By following this ideology, journalism functions as a watchdog of society, providing relevant information – impartial, fair, credible, and independent [3]. However, newsrooms are facing constraints in terms of human resources and budget cuts [23]. This phenomenon is coupled with the increase in input for news: previously non-public data is to an increasing degree published openly in a digital format. In this new environment, automation of news production is becoming an attractive option for newsrooms struggling to meet the demands of delivering relevant news at different localization and personalization levels.

In order to support news media companies in producing news from the increasing amounts of data, this paper investigates how news can be automatically extracted from open structured data sources. More specifically, we investigate the automated detection of newsworthiness from structured data using statistical analysis and how automating the generation process can provide added value to the public. This investigation is conducted using a real news *natural language generation* (NLG) system as a case study. The research is guided by first understanding the journalistic process and the steps journalists go through, from analyzing data to producing the final story and presenting it in a format best adapted for audiences.

The work presented in this paper shows the feasibility of automatically detecting what is newsworthy from structured data. In

addition, the paper illustrates how the automation of news generation can provide added value to audiences by expressing and presenting news to readers in multiple languages and in an interactive format allowing for localization and personalization, increasing the potential for media companies to grow their audiences and in turn revenue.

In Section 2 we review relevant related work on news creation both from the journalistic and automation perspectives. Section 3 presents the research problem guiding the study and describes the context of the NLG system case study. Sections 4 details our proposed approach for automatically finding newsworthy facts or events in an automated news generation system and Section 5 describes our case study system and how such it can add value for audiences. Finally, in Section 6 we examine our findings with respect to the research problem.

2 BACKGROUND AND RELATED WORK

We review work in the area of journalistic work, what is considered news, and how it is created. We are motivated by first understanding the manual news creation process, before attempting to automate it. Then we look at how the news generation process can be automated, considering in particular research on identifying newsworthiness automatically from structured data.

2.1 News Creation in the Newsroom

2.1.1 Role of Journalism. As we intend to create automated processes that can fit into the existing norms and values of news media, we must identify how the journalistic field of news production functions.

The creation of *news* appears to follow a similar set of characteristics around the globe, with some variation, based on a occupational ideology defining the profession of journalism [11]. This ideology, seen here as a system of beliefs, is rooted in a set of values, formal codes and strategic rituals that create the sphere of understanding of what constitutes journalism. The values and codes defining journalism as a profession can be condensed into five categories: public service, objectivity, autonomy, ethics, and immediacy [11, 20]. By following these ideals, journalism is acting as a watchdog of society, providing relevant information – impartial, credible, and independent – produced with immediacy within a framework of ethics and legitimacy [11].

The traditional journalistic work-process in news media production is characterized by routinized and repeated practices of identification, selection, and gathering of information [2, 35]. This information is then modified and shaped according to culturally relevant norms of journalistic storytelling. Norms and practices, such as *gatekeeping*, an innate form of "winnowing down a larger number of potential messages to a few" [35], based on both organizational and individual tacit knowledge of news selection, and the *objectivity norm*, a strategic ritual for presenting facts in order to uphold authority [38], have been formed as responses to handling the vast amounts of material available in a fast-paced environment. Given the realities of producing news with limited resources and aims for profit, media companies apply varying methods for identifying what the audience wants, what the organization is capable of producing, and what sources for stories are available [35]. As such, the journalistic routines aid in the daily work of news creation [2], particularly in an era where there are less hands on deck, increasing amounts of data and information available, including comments and feedback from audiences.

2.1.2 Defining News, News Values, and Newsworthiness. As humans, we have a natural curiosity for information about our surroundings that is new to us – stories that might provide relevant information for decision-making, strategic benefit, or social clout [5]. News gives us meaningful knowledge of events that we might not have experienced ourselves, and it functions as a valuable asset to be shared with others. Transferring this inherent human knowledge onto automated news creation requires identification of the attributes that transform information into stories.

Technological advances during the last century have brought on significant changes to how news is produced [32]. These advances, however, have not radically altered what journalists view as newsworthy [10]. News is not a direct reflection of reality, but a modified and curated version of selected events [2]. It has been argued that there are no given events "out there" that are news in themselves [29], but rather circumstances that are elevated to newsworthy status and constructed as news. What becomes news is in part defined by a vast, fuzzy list of news values, or news criteria. The newsworthiness of an event is most often defined by examining its inherent ability in satisfying one or more of the requirements for a news item.

The most influential study of news values, presented by Galtung and Ruge in 1965 [12], theorized that the clearer and easier to understand a story was, the more likely it was to be selected for publication. They outlined 12 factors for news selection; *frequency*: events that fit the publishing schedule are more likely to be picked up than long processes; *threshold*: the intensity and impact of an event affect its value; *unambiguity*: clear and easily understood stories are more popular; *meaningfulness*: the cultural closeness to the audience; *consonance*: the predictability of the event can elevate its value (e.g. elections and summits); *unexpectedness*: rare but interesting events are likely to be selected; *continuity*: a familiar story already ongoing in the news is likely to be followed up; *composition*: an event can be included in the broadcast or newspaper if it fits the overall outline, even if its innate news value is lacking; *reference to elite nations, elite people and persons*: events connected to a person are more likely to be brought up than mere abstractions; *reference to something negative*: bad events are interesting [12].

As society – and the way we consume news – evolves, so do the values we use in assigning newsworthiness. Harcup and O'Neill [16, 31] have in their studies looked at the different approaches to news values that have developed during the last decades, and added criteria that are relevant today. As an update of Galtung and Ruge's list of values, they additionally highlight exclusivity, conflict, audiovisuality, shareability, entertainment, drama, good news and the agenda of the news organization as factors influencing the criteria of selection [16].

Observing the newsworthiness factors presented above, we note that some of them cover similar aspects that can be grouped together. We view them as effectively creating three distinct larger categories of *topicality*, *outlierness*, and *interestingness*. Finally, an

element of subjectivity or *personalization* might affect newsworthiness on a personal level.

The criteria for news are thus not based on absolute boundaries, but are perpetually negotiated and defined by mutual agreements and judgments [39]. Increasingly, newsrooms function both as *publishers* of unique, comprehensive news stories, and *publicists* of events and stories already circulating in the media sphere [6].

2.1.3 Adding Value for News Audiences. What the audience sees and defines as news is highly relevant when considering introducing automated news generation. Traditionally, news stories are often presented in the form of an inverted pyramid – the most crucial information at the top, with additional information towards the bottom [35, 38]. The form was born out of necessity when the telegraph was invented, as the most important facts had to be relayed over first, in the case that the transmission would falter [39].

For the same reason, classic news reports are often constructed of short sentences and words. The news agency Reuters instructs their journalists to "get to the point" early on in their stories, answering the so called "5 W's" as soon as possible: who, what, when, where, and why [34]. From there on the typical news story is created by sequencing blocks of related information in a manner that moves the storytelling logically forward. A central semantic feature of a news report is that it describes an occurrence as a string of events, following a commonly understood narrative [7]. However, the limitless space of the digital era is changing these norms, as competing for visibility and "hooking" the reader are as important as ever.

2.2 Automatic Generation of News

Newsrooms place their "news nets" in places where the chance for finding "fish" is considered best [39]. Whereas traditional media outlets have to opt for a loose-holed net focusing on "big fish" instead of a catch-all blanket, automated news production has no such constraints. For instance, the US-based news bureau Associated Press has been able to increase the number of corporate earning reports 12-fold by automating the processes, producing over 3700 stories per quarter [25]. By doing so, the company increased its scope, possible audiences, and freed up time for the reporters to focus on investigative, complex, stories [25]. Automation enables the "small fish" to become as important as the "big fish" [39], providing a larger buffet of news. Other media companies such as the New Yorker, LA Times, and Forbes have also invested in the automatic production of news [7], an application area of natural language generation (NLG).

NLG refers to the automatic planning and generation of multi-sentence language by computer [18]. This normally involves three processes: deciding what to say (content determination), how to organise it (sentence planning), and how to express it (surface realization) [33]. NLG techniques have been applied to automatically generate news in various domains such as weather reporting [9, 13, 36], finance [1, 27, 30], and sports [4, 37, 40]. One common factor for all these domains is the abundance of structured data. However, the academic sources often provide relatively little information on the exact processes used to determine the newsworthy data points. When this information is available, the processes tend to be based on relatively static rules. At the same time, existing news generation systems used by commercial media houses are often closed source and their internal workings are unknown. Based on available information, it appears that many of the systems make use of story-level templates where specific data points can be placed. Thus, detection of newsworthiness is largely pre-determined by journalists writing the templates rather than the system itself.

NLG systems generally range from the classical type of multi-staged architecture described by Reiter and Dale [33] to end-to-end architectures, such as recurrent neural networks [19, 24, 28, 42]. Between these extremes, some systems combine both methods by using statistical learning methods embedded within the more classical multi-part pipeline architecture.

Particularly in the context of news production, template-based approaches are preferable over end-to-end black box architectures – such as neural networks – for several reasons. First, templates – pre-written strings of text into which variable content in embedded – produce text with known and accurate meanings in a predictable and transparent fashion. This is an absolute requirement in the news context where the text must accurately convey the underlying facts of the story. Second, the content is easily controllable, and errors can typically be located and fixed with relative ease by adding special cases. While statistical end-to-end methods have the benefit of automatically learning to model languages and to produce text, as black box models the above aspects are currently practically impossible to handle. In particular, guaranteed correctness is crucial in the domain of news, where an accuracy of only 99% is unacceptable.

Due to these requirements, a fully black-box model is not practical in the news generation context. Therefore, more classical NLG methods that make use of hand-written templates are currently the best option in automatically generating news for audiences.

For such template-based methods, the length and complexity of the template is a free variable. Systems can use very long templates that span a whole news article. Alternatively, a template-based system can use very short sentence-level templates that are then composed into longer stories. While the story-level templates allow for more fluent and better structured output, they are rigid and unable to adapt to unanticipated newsworthy facts. Moreover, they can be very time consuming to write.

2.3 Automatic Detection of News from Data

Stories can be found in data, in what is often referred to as data journalism [14]. NLG systems capable of generating news should be able to take as input non-linguistic input data, typically numerical or categorical, analyze it, interpret it, and decide how to communicate it in natural language. As such, awareness of the processes and practices behind news creation as described in Section 2.1 define how automated news are shaped.

Identifying what is newsworthy is a crucial step and is part of the 'content selection' phase of an NLG system pipeline. Gray et al. [14], provide a list or 'typology' of different kinds of data stories that can be created from data. These include measurements, proportions, internal and external comparisons, changes over time, league tables, analysis by categories, or associations. Various techniques have

been used in literature to select and determine the content and structure of the stories.

For example in their generation of weather news stories, Yu et al. [43] made use of Pattern Recognition, Abstraction and Interesting Pattern Selection approaches to select the patterns to convey in a story. The content selection was based on a measure of subjective and objective interestingness based on abnormality and unusualness. The order of the content was based on rules obtained via corpus analysis and expert input.

Another selection approach is to use the general heuristic that unusual (low frequency) events are more worthy of reporting than common events [26].

As noted in the above related works and in Section 2.1, a large factor in newsworthiness is something that can be characterized as outlierness or unexpectedness. That is, events that are outliers when contrasted to the set of normal, day-to-day events are often at least somewhat newsworthy. Thus, research from the fields of outlier detection and novelty detection is highly relevant. Many surveys into these extremely large fields have been published [8, 15, 17], and therefore we will only provide a brief overview of the most relevant details.

Outlier or anomaly detection methods can be divided into three categories: supervised methods, semi-supervised methods, and unsupervised methods [8]. This categorization is based on the types of training data required by the detection method. In supervised learning, a model is trained based on known samples of both normal and anomalous types. In semi-supervised learning, only samples of normal data are used to build the model. Finally, in unsupervised methods no labeled example data is required.

The assumptions related to these categories differ distinctly. Supervised methods assume that some process has previously labeled large amounts of training data [8]. Such processes are often manual and thus slow and expensive. This, in turn, causes supervised detection methods to be difficult to transfer to new domains. At the same time, unsupervised methods assume that the anomalous data points are significantly more rare than the normal data points [8]. This assumption can be reasonable in many cases, but is not guaranteed to be correct. The semi-supervised methods fall somewhere between these two extremes. In the context of news generation, where a system's transferability between domains is key, taking into consideration the cost of labeling training data for supervised and semi-supervised methods, we find unsupervised approaches preferable.

3 RESEARCH DESIGN

3.1 Research Problem

Our main objective is to investigate how automation can support media companies and journalists in their work of producing news. Inspired by the review of the background literature, we set out to answer two primary questions. First, whether the journalistic process of newsworthiness determination can be automated. As this question is very broad, we limit the scope to structured data, thus ignoring the difficulties associated with fields such as natural language understanding. Second, we wish to understand how to best use the findings of a newsworthiness determination process to provide value for audiences. More concretely, we investigate how newsworthiness can automatically be identified in structured data, as well as, how then the newsworthy data can be expressed as text. This investigation is conducted in the context of a case study, described in the next section.

3.2 Context of the Case Study

The case study presented here is a practical NLG system that produced news coverage of the results of the Finnish Municipal elections that took place in April 2017. In Finland, municipal elections take place every four years and are a large political event. The election results were publicly released immediately after the completion of the first count by the the Finnish Ministry of Justice. They were provided online in a machine-readable structured format, and were thus in the ideal format to act as input for a NLG system.

The results files included the results for each party on the level of the whole country, each of the 13 electoral districts, 311 municipalities, and 2,012 polling stations. For each of the 33,316 candidates, the data included details of their success in their own municipality (each candidate is only votable in a single municipality) and all of the municipality's polling stations.

In total, 725,066 valid combinations of locations and entities (party, candidate, or none) were possible. Each of these combinations corresponded to a distinct topic for a news story, thus providing the ability to produce news at varying levels of locality. In terms of locality, the system was developed to generate news articles in three languages, i.e., Finnish, Swedish, and English. The Finnish context calls for multilinguality as Finland is bilingual, with both Finnish and Swedish being official languages. With international audiences in mind – and as a catch-all for other language minorities – it was also desirable to produce articles in English.

4 PROGRAMMATIC DETECTION OF NEWSWORTHINESS

The journalistic process must start with determining which events are newsworthy. Newsworthiness is a fuzzy and subjective phenomenon, and as such, any algorithm for defining newsworthiness is going to be somewhat lacking in comparison with the inherent knowledge of news values present in a newsroom. On the other hand, an explicit representation of newsworthiness criteria to a computer system has the benefit of transparency: the criteria can be discussed, debated, and adjusted, and the system's decisions can be explained.

4.1 Modeling newsworthiness

We propose an approach where the newsworthiness of an event is represented by a factor for each of the four categories identified in Section 2.1.2: topicality, outlierness, interestingness, and personalization. These factors are non-negative real numbers where value 0 means 'absolutely not newsworthy' and value 1 is a neutral value; otherwise the factors represent relative newsworthiness between events. For instance, value 2 means that the event is twice as newsworthy as a neutral event. The product of these factors then determines the overall newsworthiness of the event.

Topicality is a function of the age of the events being described – older events are less topical – and whether the events fit into the current public discourse. It corresponds to things like continuity

and composition. For example, a rare event might make the previous instance of the same event topical again, even if a significant amount of time has passed since said previous instance. We therefore model topicality as $T(e, d)$ where e is the event or fact we consider and d is the public discourse and context. The domain of this function is $\mathbb{R}^+ \cup \{0\}$: while the concept of "completely untopical" is sensible, there is no clear definition for "maximum topicality".

Outlierness, on the other hand, refers to factors like unexpectedness. It is something that is measured in terms of some context. Thus, the outlierness value of some event is clearly a function of that event and the set of "relevant events", where relevancy is some certain frame of reference. For example, in the context of the Finnish municipality election news (see Section 3.2), the fact that some candidate received 5,321 votes can be a significant news item in a small community with 10,000 voters while being complete trivia in the case of the capital city, with hundreds of thousands of voters. As such, we model outlierness as $O(e, E_r)$, where e is the event or fact we consider and E_r is the set of all relevant events, such as vote counts of other candidates in the same municipality, with $e \in E_r$. Similarly to topicality, the domain of $O(\cdot)$ is $\mathbb{R}^+ \cup \{0\}$: while "not an outlier" is a reasonable interpretation for zero, the concept of "maximum outlier" is not clearly defined.

Next, we consider interestingness as the intrinsic newsworthiness related to the people and places associated with the event. In the context of elections, for instance, results on the municipal level are more intrinsically interesting than the results of an individual polling station. At the same time, if a candidate was somehow famous otherwise, that fame might make their results more interesting. In other words, things that happen to interesting people, in interesting places can be newsworthy even if the event itself is mundane. Thus, we model interestingness as $I(e)$, where the event or fact e contains the information on what entities (people, locations, etc.) are related to said event or fact. As with the previous factors, and for similar reasons, the domain of $I(\cdot)$ is $\mathbb{R}^+ \cup \{0\}$.

Finally, the user's subjective view is modeled as a personalization factor $P(e)$ with a domain of non-negative real numbers. The distinction between P and I is in the perspective: $I(\cdot)$ refers to the general interestingness of the fact and is essentially the journalist's view on the importance of the fact. At the same time, $P(\cdot)$ refers to the personal interest of an individual reader. This distinction is important in cases where some fact or event is relatively unimportant in general, but is of high interest to some individual user. For example, while the election results of some polling station are in general uninteresting, they might have a large personal interest factor for those voters that voted at that polling station.

Formulated in the above terms, newsworthiness can then be modeled as

$$N(e, d, E_r) = T(e, d) \times O(e, E_r) \times I(e) \times P(e) \tag{1}$$

In terms of computing the newsworthiness, determining the event e is trivial. The set of relevant events E_r is slightly more complex, as "relevant" is context-dependent. For example, if e is the number of votes a certain candidate received in a certain geographic location in the municipal elections, we consider as related all election results that pertain to either the same candidate or the same geographic location. The implication of this will be discussed later.

Determining d – the public discourse at the time – is still a complicated task for an automated system. The system would be required to follow and more crucially understand the public discourse, a feat far from trivial. As such, we do not delve into that part of topicality, focusing solely on the temporal aspect. In cases where the system is real-time or near real-time, the temporal aspect can always be assumed to be such that the topicality is of some reasonable value. In cases where the system is not real time, the topicality can then be determined by a human actor that decides whether to run the system or not. This effectively reduces the topicality factor into a constant for most applications. Thus, the model for newsworthiness becomes

$$N(e, E_r) = O(e, E_r) \times I(e) \times P(e) \tag{2}$$

Next, we will discuss an instantiation of this generic method for determining newsworthiness in the context of the NLG system created for the 2017 municipal elections in Finland.

4.2 Measuring newsworthiness in election results

Following the model expressed in equation 2, we determined those events and facts from the municipality election data that were newsworthy (see Section 3).

We modeled the tabular data provided by the Ministry of Justice using a sextuplet schema (`entity, entity_type, location, location_type, value, value_type`). In this schema, `entity` and `entity_type` together uniquely describe the party or candidate the data point pertains to. Here, `entity_type` describes whether the entity is a candidate or a party and `entity` is an identifier unique within the type. Similarly, `location` and `location_type` uniquely describe the geographic location to which the data point pertains to. The possible `location_type` values are the whole country, an electoral district, a municipality and a polling station in decreasing order of size. Finally, `value` and `value_type` describe a numeric or categorical value and its meaning. For example, a `value_type` might be the string "*total_votes*" and the value would the be a number describing the number of votes some entity received in some geographic location.

4.2.1 Outlierness determination. For determining the outlierness of a value, e.g., number of votes, we use a method that is based on Inter Quartile Ranges (IQR) (see for example [8]). The basic IQR test simply labels as anomalous all observations that are more than $1.5 \times (Q_3 - Q_1)$ away from the nearer of Q_1 or Q_3, which are the first and the third quartiles of the observed data, respectively [8]. While this method is unable to properly handle multidimensional distributions outside of some special cases, it provides a simple and non-parametric outlier detection method for univariate data.

We are interested in ranking values based on how outlying they are, so we modified the calculation to produce continuous values $O(e, E_r)$ of outlierness as follows:

$$O(e, E_r) = \frac{|e - Q_{2,E_r}|}{Q_{3,E_r} - Q_{1,E_r}}. \tag{3}$$

Here, e is assumed to be a numeric value, E_r to be a collection of such numeric values. Q_{i,E_r} refers to the i-th quartile of E_r.

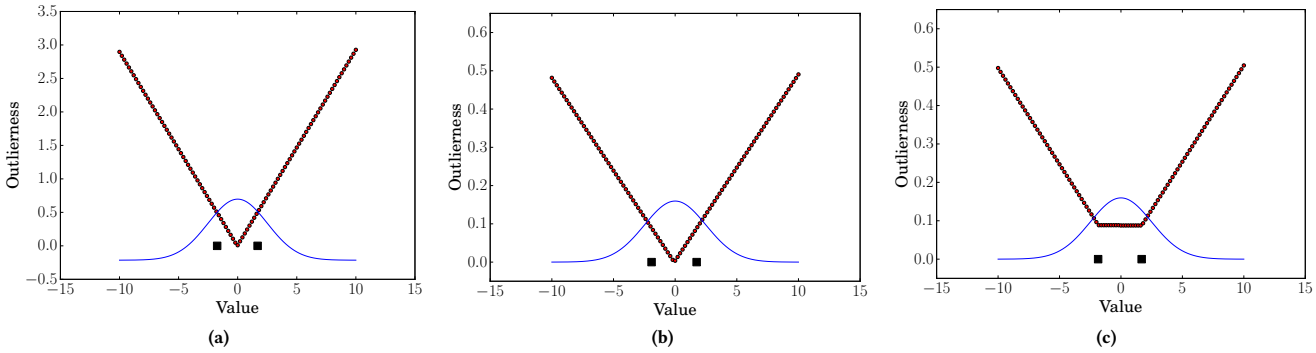

Figure 1: Behavior of different versions of the outlierness metric: (a) initial version, (b) weighted version, (c) weighted and flattened variant. Data points are normally distributed, illustrated by the thin (blue) line. The two black squares indicate the first and third quartile of the distribution. Outlierness scores are shown with thick (red) lines, with their values on the Y-axis.

The behavior of this function for data E_r that is normally distributed in $\mathcal{N}(0, 2.5)$ is illustrated in Figure 1a. The thin (blue) line shows the distribution of the data along the X-axis, and the thick (red) line gives the corresponding outlierness scores. The black squares indicate the first and third quartiles, used in the computation of $O(e, E_r)$ in Equation 3. As can be seen, the scores now allow comparing how outlying different values are.

While the above outlierness score has the nice property that it is distribution-driven, it is at the same time sensitive to the sample size in the sense that a larger data set will produce more outliers than a smaller one with the exact same distribution. We wish to reduce this effect, and we weight this initial outlierness value down by a factor related to the size of the set of data points:

$$O(e, E_r) = \frac{|e - Q_{2,E_r}|}{Q_{3,E_r} - Q_{1,E_r}} \times \sqrt{\frac{1}{\sqrt{|E_r|}}}. \tag{4}$$

Figure 1b illustrates the behavior when the number of data points is $|E_r| = 1000$. (Note the change in the scale of the Y-axis; the thin (blue) distribution stays identical but the outlierness scores are scaled down.)

Next, we observe that data points between the first and third quartiles Q_{1,E_r} and Q_{3,E_r}, i.e. data points that are not outliers, also receive varying scores of outlierness. With respect to outlierness, however, these data points should be considered mutually equal. We thus modify the calculation as follows:

$$O(e, E_r) = \begin{cases} min_{i \in \{1,3\}} O(Q_{i,E_r}, E_r) & \text{if } Q_{1,E_r} < e < Q_{3,E_r} \\ \frac{|e - Q_{2,E_r}|}{Q_{3,E_r} - Q_{1,E_r}} \times \sqrt{\frac{1}{\sqrt{|E_r|}}} & \text{otherwise} \end{cases} \tag{5}$$

This modified version assigns for values between the first and third quartiles the lower of the two outlierness value of the first or third quartile. The modification results in the behavior presented in Figure 1c.

Finally, we note that the outlier calculation is undefined in those reduced cases where more than half of the data points are identical and $Q_{1,E_r} = Q_{2,E_r} = Q_{3,E_r}$. In this special case we use the following alternative formulation, $O'(e, E_r)$, that handles separately the two tails and the bulk of data points identical to $Q_{1,E_r} = Q_{2,E_r} = Q_{3,E_r}$:

$$O'(e, E_r) = \begin{cases} C_1 \times \frac{e - Q_{2,E_r}}{\max E_r - Q_{2,E_r}} \times \sqrt{\frac{1}{\sqrt{|E_r|}}} & \text{if } e > Q_{2,E_r} \\ C_2 \times \sqrt{\frac{1}{\sqrt{|E_r|}}} & \text{if } e = Q_{2,E_r} \\ C_1 \times \frac{Q_{2,E_r} - e}{Q_{2,E_r} - \min E_r} \times \sqrt{\frac{1}{\sqrt{|E_r|}}} & \text{if } e < Q_{2,E_r} \end{cases} \tag{6}$$

Here, the constants C_1 and C_2 are free variables. We selected for our case study $C_1 = 2$ and $C_2 = 0.5$ based on empirical tests, i.e, the bulk has outlierness value 0.5 and any other data points have value 2.

4.2.2 Interestingness detection. For determining the interestingness value, we make use of the individual facts as sextuplets (`location`, `location_type`, `entity`, `entity_type`, `value`, `value_type`). The `X_type` components allow us to make generalizations over groups of items. As a concrete example, the fact that the National Coalition (NC) party got 25 seats in the new council of the capital city Helsinki (Municipality number 91) would be represented as the sextuplet (`91`, `municipality`, `NC`, `party`, `25`, `seats`). From this representation, the interestingness evaluation function observes only the `location_type` ("municipality") , the `entity_type` ("party") and the `value_type` ("seats"). In other words, the actual values are ignored. We then build the interestingness value as

$$I(e) = I'(e_{location\_type}) \times I'(e_{entity\_type}) \times I'(e_{value\_type}) \tag{7}$$

where I' is a function that returns a real-valued weight based on the input, and $e_{variable}$ indicates the *variable* of the sextuplet e. These weights are encoded manually by the journalist in a separate configuration file. As an example, we settled on the following $e_{location\_type}$ weights

$$I'(e_{location\_type}) = \begin{cases} 1.0 & \text{if } e_{location\_type} = \text{`}municipality\text{'} \\ 0.7 & \text{if } e_{location\_type} = \text{`}country\text{'} \\ 0.5 & \text{otherwise} \end{cases} \tag{8}$$

for the municipal election application through empirical testing.

Similar functions were implemented for the `entity_type` and `value_type` interestingness calculations. As they are relatively long

are highly specific to the problem domain, they are not presented here. It is of note, however, that especially the interestingness value of `value_type` is again a combination of multiple factors, such as "does this `value_type` describe a change in value" and "does this `value_type` describe something related to the number of seats", and so forth.

4.2.3 Personalization factor. In our case study, P is not explicitly defined. Rather, we surrender control over selecting the focus of the generated article to the user by allowing them to fix the `location` and `location_type` values and optionally the `entity` and `entity_type` values. This can be seen as a function for $P(\cdot)$ that returns some non-zero constant for the selected values and zero otherwise.

5 A CASE STUDY IN GENERATING TEXTUAL NEWS ARTICLES

The underlying reason for detecting newsworthy facts from structured data is to create value for audiences. We believe that automatically generating news stories is an excellent way of providing value to external customers, the in-house journalists. Thus, having determined the newsworthy events and facts as detailed in Section 4, they must next be expressed to the readers in a suitable format.

Non-natural language expression methods such as graphs or tables, although useful, are unable to express the full range of information available to textual expression. Furthermore, some studies have shown that textual representations reduce visual complexity and mental workload and enhance decision-making in comparison to graphs [21, 41], thus suggesting that NLG might offer a viable approach in presenting real, complex and dynamic data sets [41]. In addition, by first producing textual output, it is possible in future to be able to expand the NLG system to also produce audio and or video output.

We present next some details of the cases study system that creates natural language news articles within the election results domain as well as the approaches taken to structure stories based on identified newsworthy facts.

The election news generation system was built as a website. A screen capture of the user interface is presented as Figure 2. Using searchable drop-down menus, the users of the website can direct the system towards producing a news article about some focus they are interested in. The users can select the language (Finnish, Swedish, or English) and the geographic focus of the news article. They can also select to read an article on the election results in general, or alternatively focus the article on an individual party or candidate within the selected geographic focus. In cases where the selected focus is extremely narrow, for example if an individual candidate of which little data is available is selected, the system automatically expands the article to include content about other entities in the same geographic area. In such cases, the first paragraph will always be about the user's original selection and the expanded content is displayed only after the first paragraph. In our case study a total of 725,066 articles with unique focuses are readable for each of the three languages.

The case study system is publicly available at www.vaalibotti.fi. The online setting allows for some user-interaction as well as sharing of news articles with friends. The system also recommends additional news articles to the user.

Figure 2: The user interface of the implemented system. The user has requested an English language article about the election results in the municipality of Jyväskylä without specifying any other focus, such as a specific party or candidate. All other decisions regarding the content and structure of the story were made by the system.

5.1 Structuring Stories Based on Newsworthiness

In the structuring of news articles, we wish for the automated system to always report the most newsworthy facts first, depending on the specific user request. At the same time, we do not wish to restrict the story structures in advance to a small set of pre-written, large story-level templates: the system should produce text based on what is most newsworthy, not based on what was is anticipated as being the most newsworthy. We use smaller sentence-length templates, i.e., ones averaging six to eight words in English. This way the system can always choose to start the first paragraph with the fact that was calculated as being most newsworthy. This is especially important when the system produces highly tailored content specific to each user request.

When a user selects a focus (i.e., location and or entity), the system produces an article by first determining the data points pertaining to the user's selection that are most newsworthy, then selects the templates that most appropriately fit those data points. This reveals a crucial downside to using smaller templates: the decline in the fluency of the story. If content is selected and ordered purely based on what is newsworthy, the output of the system will likely be a text that very quickly jumps from one subtopic to another without regard for any larger structure. This phenomena can be combated by constructing the story in paragraphs, where each paragraph is limited to one subtopic (in the case study system, to a single candidate or party) and changes in subtopic are restricted

```
fi: {who} saa [{location, case=ssa}] enemmistön valtuuston paikoista
en: {who} secured a majority of the seats [in {location}]
| percentage_seats > 50.0, place_type = municipality

fi: {who} sai [{location, case=ssa}] {value, abs} ääntä vähemmän kuin edellisissä kuntavaaleissa
fi: {who} menetti [{location, case=ssa}] {value, abs} ääntä edellisistä kuntavaaleista
en: {who} dropped {value, abs} votes since the last municipal election [in {location}]
en: {who} got {value, abs} fewer votes than in the previous municipal election [in {location}]
| total_votes_change < 0
```

Figure 3: An example of the templating language used for the municipal election application. Lines starting with 'fi:' and 'en:' are templating strings in the indicated language (Finnish and English, respectively) and lines starting with a vertical bar define constraints for the groups of templates. Groups are separated by empty lines. In the template strings, sections in square brackets are optional and sections in curly brackets are replaced by values from the sextuplets.

to happen only between paragraphs. In other words, the newsworthiness criteria of composition (see Section 2.1) also applies within individual news stories, when deciding which facts to report.

The fluency of output can be further improved by aggregating sentences to reduce redundancy. Using the election system as an example, the system combines the smaller sentences "Party A got X votes" and "Party A got Y seats" to the larger and more fluent sentence "Party A got X votes and Y seats".

To further improve the fluency of the output, the system employs entity name resolution techniques to avoid repetition. For example, if the the name of a party or a candidate is mentioned twice in a row, the second mention can be replaced with a pronoun. Similarly, once a previously mentioned party or candidate is referenced later in the article, with some other entity referenced between these instances, a shorter form of the candidate's or the party's name can be used.

As an example, the story shown in Figure 2 was generated based solely on the user request "the language should be English" and "the story should discuss the municipality of Jyväskylä". All other decisions regarding the content and structure of the story were made by the system. The technical details of this process are explained in more detail in [22].

To allow relatively non-technical journalists to contribute to the creation of these templates, a custom templating language was created. This language consists of template strings such as "`{entity} got {value} votes in {location}`" and constraints that specify, for example, that the previous template can only be used with sextuplets where the `value_type` is "votes". An excerpt of the templating language is presented as Figure 3.

Employing this templating language provides several benefits. First, as mentioned above, it allows relatively non-technical journalists to efficiently write templates with complex constraints with minimal help from programmers and other technical staff. Secondly, adding new languages to the template set does not require creation or modification of the constraints but mere translation of the template text. Thus, the approach allows news to be made available to audiences in several languages. This multilinguality stands in contrasts to how normal news production works, with individual journalists usually only creating content in one language.

More details of language generation within the case study system are presented in [22].

6 DISCUSSION

With the aim of supporting the journalistic process, this paper has presented an approach for automatically identifying newsworthiness from structured data and illustrated how automation of news production can add value for audiences by automatically generating news. In the next subsections, we discuss the research problem outlined in Section 3 with respect to the findings and related works.

6.1 Identifying Newsworthiness from Structured Data

The production of news by journalists is driven by first identifying what events are newsworthy. A set of news values or criteria usually guides this process (see Section 2.1), and in this paper we investigated approaches that can be used to automate the identification of newsworthy events from structured data. We identified a set of criteria that encompass the set of news values, i.e., topicality, outlierness, interestingness, and personalization. Furthermore we proposed that newsworthiness can be modeled as a product of all these four criteria, where we provided formulations for determining each criterion. In addition, we illustrated its implementation in identifying newsworthy facts from structured election result data.

Automating this task results in several advantages in terms of breadth of coverage and transparency in the process. The automation of finding news in the era of increased availability of data sources ensures that media companies efficiently make use of the data and report even on the "small fish" news which they would not have had the time or resources to report. This means that they are able to cover news that is of interest to even local or small audiences, thus improving business offering and providing added value to audiences. Of course, what is newsworthy or of interest to audiences is heavily dependent on the data source. In our example demonstration, we only looked at newsworthiness from one source, but the model can be applied to other data sources, evaluating them on the four criteria.

In addition, the formulation for identifying what is newsworthy is very transparent in comparison to the fuzzy selection journalists use (described in Section 2.1), although human journalists have a better world view and understanding and thus have multiple sources of information for identifying what is newsworthy.

6.2 Adding Value for Audiences

In this work, we have observed that the main benefits of an automated system for audiences resides in the ability to generate large amounts of tailored content, in particular, the ability for the reader to interact and tailor the content based on their own interests. Humans have a natural curiosity for information about surroundings and events that are new – stories that might provide relevant information for decision-making, strategic benefit, or social clout [5]. Therefore, the ability to select or narrow down to news that is of personal interest or relevancy is valuable.

Traditionally, the practice of *gatekeeping* as covered in Section 2.1, has been left in the hands of the journalists, whereby they apply methods for identifying what the audience wants, what the media company is capable of producing, and what sources for stories are available [35]. However, this results in a limited selection of news being available to the public. With automating the generation process, there is no need for winnowing down news. The system can produce as many news articles as possible and put the power of selection in the reader's hands. In the context of the election NLG system, this was demonstrated by having the system have the capability to produce a high quantity of news articles (over 700,000 in each of the three languages), and then allowing the audience to narrow down to those articles they want to read by selecting a party or candidate in a specific location. Suggestions for related news can be provided to provide support to the reader to find even more news that might be of interest to them. Future versions can be implemented to allow more tailoring by allowing the reader to for instance say that they want to read news articles only about parties or only about candidates.

An advantage of this automation approach is that it further frees newsrooms from focusing only on the big stories. They can thus cast a big net to catch even the small news, which in turn can help reach a broader audience and following. When generating news from structured data, a natural limitation is that these nets can only be cast to where structured data exists.

In addition, due to a lack of resources in newsrooms and increased data availability, potentially interesting stories might end up not being written. In such cases, the newsrooms are unable to reach some audiences. This is immediately apparent when one considers the traditional unilingual newsrooms. There, all audiences in minority languages – no matter how local – are unreachable. In the case study system, news articles were generated in three languages. These three languages ensure that almost every Finnish person was reachable through one of the languages. Such a similar multilingual capability is only achievable in manual news production through considerable effort.

This work has thus identified that through the automation of news generation, value can effectively be provided to audiences. This is achieved by producing a large number of news articles that the newsrooms are unable to manually produce and by targeting demographic segments that are too small to sustainably target with manual news production.

6.3 Determining Relatedness

While the algorithm we present for determining newsworthiness is in principle simple, questions regarding relevancy still remain. In our case study system, the user manually specified a geographic location and was then presented with a news-article that was specific to that geographic location. This approach works well in many cases, but has a shortcoming in being unable to present highly interesting and related news from – for example – the neighboring municipality. The problem, in other words, is that in some cases adding tangentially related information to an article is the right thing to do, and sometimes it is not.

We have taken first steps in implementing a fuzzy matching system that allows a news article to contain facts pertaining to "relevant" geographic areas if the newsworthiness of those facts is high. To limit the extent to which these tangentially related facts are included, the system increasingly penalizes facts from geographic locations that are further away from the user's original selection in either size or location. We have, as of yet, been unable to generate a system for determining and weighing relatedness that works sufficiently well to match our human intuitions.

A further complication, in most of our attempts, is that if some fact is sufficiently newsworthy, it will be presented in every news article, irrespective of the original focus. This becomes jarring after only two or three articles are read. The most obvious solution to this problem is recording a discourse history and down-weighting the newsworthiness values of facts that have already been presented to the user multiple times. This, however, would require per-user tailoring of the content and as such presents additional challenges.

6.4 Limitations and Future Work

We acknowledge that the external validity of the study is compromised by the fact that our case study focuses on a single domain. However, we expect the newsworthiness formulation and approaches for creating value, that is, multilinguality, interactivity and localization, apply to other domains as well. In addition, we focused our automation efforts on the news production aspect of journalism, thus ignoring large sectors of the vast journalistic field, encompassing aspects of public debate and civic discourse. More work is required to better understand how these other aspects of journalism are related to automation efforts.

In future work, we intend to test the presented approaches by transferring them to new domains. Potential improvements during this process include broadening the concept of interactivity, from reader - content interaction (for instance allowing readers to affect the weights used to assign interestingness to facts) to also newsroom - reader interaction.

In this work, we represented the identified newsworthy facts as text because of its explanatory power. As future work we intend to further investigate whether including graphical illustrations would provide added value to some audiences.

7 CONCLUSIONS

In this paper we have presented approaches for automatically finding newsworthy events from structured data as well as approaches for providing value for news readers. Based on these approaches, we demonstrated their implementation in a real NLG system that generated news articles based on election data. Our work with automated journalism indicates that the algorithmic detection of novel newsworthy content from structured data is feasible and it

fulfills the purpose of supporting journalists and media companies to produce a large quantity of news articles that serve a wide audience at varying local levels and multiple languages. Furthermore, automated news generation can bring the audience closer by providing interactivity with the generation system, thus producing added value for news customers.

ACKNOWLEDGMENTS

This work has been funded in part by the Finnish Funding Agency for Innovation (Tekes) and by the Academy of Finland under grant 276897 (CLiC).

REFERENCES

[1] Peggy M Andersen, Philip J Hayes, Alison K Huettner, Linda M Schmandt, Irene B Nirenburg, and Steven P Weinstein. 1992. Automatic extraction of facts from press releases to generate news stories. In *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 170–177.

[2] Lee B Becker and Tudor Vlad. 2009. News organizations and routines. *The handbook of journalism studies* (2009), 59–72.

[3] Daniel A Berkowitz. 2009. Reporters and their sources. *The handbook of journalism studies* (2009), 102–115.

[4] Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, and Leo Wanner. 2012. Perspective-oriented Generation of Football Match Summaries: Old Tasks, New Challenges. *ACM Trans. Speech Lang. Process.* 9, 2 (2012), 1–31.

[5] Brian Boyd. 2009. *On the origin of stories.* Harvard University Press.

[6] Axel Bruns. 2008. 3.1. The Active Audience: Transforming Journalism from Gatekeeping to Gatewatching. (2008).

[7] David Caswell and Konstantin Dörr. 2017. Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice* (2017), 1–20.

[8] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *Comput. Surveys* 41, 3 (2009), 15.

[9] Shyi-Ming Chen and Ming-Hung Huang. 2014. Automatically generating the weather news summary based on fuzzy reasoning and ontology techniques. *Information Sciences* 279 (2014), 746–763.

[10] Simon Cottle and Mark Ashton. 1999. From BBC newsroom to BBC newscentre: On changing technology and journalist practices. *Convergence: The International Journal of Research into New Media Technologies* 5, 3 (1999), 22–43.

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

[12] Johan Galtung and Mari Holmboe Ruge. 1965. The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of peace research* 2, 1 (1965), 64–90.

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

[14] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. 2012. *The data journalism handbook: how journalists can use data to improve the news.* " O'Reilly Media, Inc.".

[15] Manish Gupta, Jing Gao, Charu C Aggarwal, and Jiawei Han. 2014. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering* 26, 9 (2014), 2250–2267.

[16] Tony Harcup and Deirdre O'Neill. 2016. What is news? *Journalism Studies* (2016).

[17] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (2004), 85–126.

[18] Eduard H Hovy. 1992. *Natural language generation.* Technical Report. DTIC Document.

[19] Chloé Kiddon, Luke Zettlemoyer, and Choi Yejin. 2016. Globally Coherent Text Generation with Neural Checklist Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 329–339.

[20] Bill Kovach and Tom Rosenstiel. 2007. *The elements of journalism: What newspeople should know and the public should expect.* Three Rivers Press (CA).

[21] Anna S Law, Yvonne Freer, Jim Hunter, Robert H Logie, Neil McIntosh, and John Quinn. 2005. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of clinical monitoring and computing* 19, 3 (2005), 183–194.

[22] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-Driven News Generation for Automated Journalism. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*. ACL. To Appear.

[23] Carl-Gustav Linden. 2017. Decades of Automation in the Newsroom: Why are there still so many jobs in journalism? *Digital Journalism* 5, 2 (2017), 123–140.

[24] Konstantin Lopyrev. 2015. Generating news headlines with recurrent neural networks. *arXiv preprint arXiv:1512.01712* (2015).

[25] Francesco Marconi and Alex Siegman. 2017. The future of automated journalism: A guide for newsrooms in the age of smart machines. (2017).

[26] Mark T Maybury. 1995. Generating summaries from event data. *Information Processing & Management* 31, 5 (1995), 735–751.

[27] Sheila Mendez-Nunez and Gracian Trivino. 2010. Combining semantic web technologies and computational theory of perceptions for text generation in financial analysis. In *Fuzzy systems (fuzz), 2010 IEEE international conference on*. IEEE, 1–8.

[28] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model.. In *Interspeech*, Vol. 2. 3.

[29] Harvey Molotch and Marilyn Lester. 1974. News as purposive behavior: On the strategic use of routine events, accidents, and scandals. *American sociological review* (1974), 101–112.

[30] Liubov Nesterenko. 2016. Building a System for Stock News Generation in Russian. In *Proc. of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, C Gardent and A Gangemi (Eds.). ACL, Stroudsburg, PA, 37–40.

[31] Deirdre O'Neill and Tony Harcup. 2009. News values and selectivity. *The handbook of journalism studies* (2009), 161–174.

[32] John Pavlik. 2000. The impact of technology on journalism. *Journalism Studies* 1, 2 (2000), 229–237.

[33] Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems.*

[34] Thomson Reuters. 2008. *Reuters Handbook of Journalism.* Thomson Reuters.

[35] Pamela Shoemaker and Stephen D Reese. 1996. Mediating the message White Plains. *NY: Longman* (1996).

[36] Somayajulu Sripada, Ehud Reiter, and Ian Davy. 2003. SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update* 6, 3 (2003), 4–10.

[37] Mariët Theune, Esther Klabbers, Jan-Roelof de Pijper, Emiel Krahmer, and Jan Odijk. 2001. From data to speech: a general approach. *Natural Language Engineering* 7, 01 (2001), 47–86.

[38] Gaye Tuchman. 1972. Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of sociology* 77, 4 (1972), 660–679.

[39] Gaye Tuchman. 1978. Making news: A study in the construction of reality. (1978).

[40] United Robots. 2017. Rosalinda for Sports. (2017). retrieved on 24th April 2017, http://www.unitedrobots.se/produkter-1/.

[41] Marian Van Der Meulen, Robert H Logie, Yvonne Freer, Cindy Sykes, Neil McIntosh, and Jim Hunter. 2010. When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. *Applied Cognitive Psychology* 24, 1 (2010), 77–89.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[43] Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. 2007. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering* 13, 01 (2007), 25–49.