A Baseline Document Planning Method for Automated Journalism

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

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

Abstract

In this work, we present a method for content selection and document planning for automated news and report generation from structured statistical data such as that offered by the European Union's statistical agency, Eurostat. The method is driven by the data and is highly topic-independent within the statistical dataset domain. As our approach is not based on machine learning, it is suitable for introducing news automation to the wide variety of domains where no training data is available. As such, it is suitable as a low-cost (in terms of implementation effort) baseline for document structuring prior to introduction of domain-specific knowledge.

1 Introduction

Automated generation of news texts from structured data – often referred to as 'automated journalism' (Graefe, 2016; Dörr, 2015; Caswell and Dörr, 2018) or 'news automation' (Linden, 2017; Sirén-Heikel et al., 2019; Dierickx, 2019) – is of great interest to various news producers. It is seen as a way of 'providing efficiency, increasing output and aiding in reallocating resources to pursue quality journalism' (Sirén-Heikel et al., 2019, p. 47). While data-to-text NLG systems are still far from common especially among the smaller, regional news industry players, at least among the larger newsrooms the use of NLG approaches has clearly been established (Fanta, 2017).

While secrecy in the industry makes it difficult to establish the commercial reality as an outsider, the limited available evidence indicates that commercial automated journalism is mostly done using rule-based methods despite a surge of academic interest in increasingly complex neural methods for NLG (e.g. Puduppully et al., 2019; Ferreira et al., 2019): Interviews of news automation users indicate that the employed methods are mostly based on templates (Sirén-Heikel et al., 2019), as are the few open source code repositories of real-world news automation systems (Yleisradio, 2018). Indeed, some NLG industry experts believe that especially end-to-end neural models do not match customer needs at this time (Reiter, 2019).

Contributing factors include a lack of control (Reiter, 2019); issues with hallucination of non-grounded output (Nie et al., 2019; Dušek et al., 2019; Reiter, 2018); the difficulty in surgically correcting any issues identified in trained neural models beyond additional training; as well as the difficulty of establishing what the 'worst case' performance of a neural model is.

In addition, we believe that that while neural NLG methods are theoretically highly transferable, the *practical* transferability of neural NLG solutions to many news domains is limited by a lack of training data. While newsrooms have extensive archives of news text, these are rarely associated with the matching data that is the 'input' for each piece of news text (E.g., MacKová and Sido, 2020, pp. 43–44, Kanerva et al., 2019, p. 247). At the same time, the non-trainable methods for NLG, too, suffer from difficulties in transferability and reusability (Linden, 2017).

In this work, we investigate document planning (selecting what content and in what order should appear in the document) for structured, statistical data-to-text NLG in the context of automated journalism targeting human journalists. We are not in search of a perfect method, but rather something that is relatively easy to implement as a subdomain-independent baseline and which can then be enhanced with domain-specific processing later-on. Such a method would make it easier to introduce automated journalism solutions to completely new subdomains within the larger statistical data domain.

2 Structuring Hard News

When queried for insight into news structure, journalists and academics often recite the concept of the "(inverted) news pyramid", where the news article is structured so that the order in which information appears in the text reflects the journalist's belief about the importance of the piece of information (Thomson et al., 2008). While the precise origin of the structure is not clear (Pöttker, 2003), it has become so prototypical that it is held self-evident in the journalistic trade literature: "*Every journalist knows how to write a traditional news text: start with the most important thing and continue until you have either said everything relevant or the space reserved for the story runs out*" (Sulopuisto, 2018, translated from Finnish).

A more rigorous analysis of the structures employed in 'hard' news is presented by White (1997), who argues that hard news articles have an 'orbital' structure consisting of a *nucleus* which represents the main point of the article and *satellites* that give context and additional information about the nucleus. White (1997) assigns the role of the nucleus to the combination of the headline and the lead paragraph of the article, and describes the subsequent paragraphs as the satellites. White (1997) identifies five possible relations between a satellite and the nucleus: elaboration, cause-and-effect, justification, contextualization and apprisal. Thomson et al. (2008), in turn, identify that the satellites can elaborate, reiterate, describe causes or consequences, contextualize or provide additional assessment. An important observation is that – as indicated by 'orbital' – these satellites are relatively freely reorderable without affecting readability or meaning. Together, these two observations indicate that a good document plan for hard news (1) prioritizes more newsworthy items and (2) contains some overarching theme (exemplified by the nucleus) so that the text as a whole is coherent, i.e. the satellites are in some way related to the nucleus.

The relations identified by White (1997) and Thomson et al. (2008) are highly similar to those identified in the more general Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), which uses similar nucleus-satellite terminology. However, whereas White (1997) and Thomson et al. (2008) analyze news text on the level of paragraphs, RST can be applied on a more fine-grained level to much shorter text spans. As RST shows that similar relations can be applied on a sub-paragraph level, we hypothesize that a reasonably approximation of a news article might be constructed by applying White's (1997) orbital theory also *within* paragraphs, by considering the first *sentence* of the paragraph a nucleus, and the others as satellites.

Importantly, we interpret the orbital theory of news structuring to suggest that – as the satellites are freely orderable – the actual *type* of relation is not as important for document planning as knowing that *some* relation exists between the satellite and the nucleus. We hypothesize that while identifying whether a specific (RST) relation exists between two arbitrary pieces of information requires domain knowledge, an approximation of whether two arbitrary pieces of information are related in *some* way could be obtained by inspecting their similarity in a domain-independent fashion.

That is, we expect that a piece of information regarding the US health care funding in 2020 is more likely to be related in *some way* to a piece of information discussing the US health care funding in 2020 than to another piece of information discussing the health care funding in Sweden in 1978. If a heuristic or similarity measure identifying such relations could be identified, it could be used together with some estimate of newsworthiness to construct paragraph and document plans that seek to maximize both the key aspects identified above: newsworthiness and the relatedness of the content.

As noted in the introduction, there is a distinction between the theoretical and the practical transferability of neural processing methods. We believe that a good baseline document planning and content selection approach should avoid the need for training data present in the many of recently proposed document planning and content selection approaches. This rules out as unsuitable most recent work that are based on learning from an aligned corpus of data and human-written texts, such as Angeli et al. (2010), Konstas and Lapata (2013), Wiseman et al. (2017), Zhang et al. (2017), Li and Wan (2018), Dou et al. (2018) and Puduppully et al. (2019).

Outside of these trainable approaches, to our knowledge, most other document planning approaches are based on '*hand-engineered*' (Konstas and Lapata, 2013), domain-specific methods. A highly relevant survey of various document planning methods is presented by Gkatzia (2016). While these previous works are – to at least some degree – domain-specific, they establish concepts

and ideas that are highly relevant for our goal. Both Hallett et al. (2006) and Gatt et al. (2009) describe a core set of information, called 'summary spine' or 'key events', that they hold as more important than the rest of the available information. They, as well as Banaee et al. (2013), also employ a numeric estimate of importance. Demir et al. (2010) identify that content already selected for inclusion in the document plan affects how well suited so-far unselected content is for inclusion. Sripada et al. (2003) identify Gricean maxims (Grice, 1975) as providing requirements for document planning and content selection.

3 Context

Our work on document planning is done in the context of a series of data-to-text NLG applications producing short highlights of structured statistical data. Importantly, the applications are intended to be deployed in contexts where they must be able to produce texts highlighting between 10 and 30 data points from datasets measured in 100.000s of data points. The resulting texts are intended to both alert journalists to potential news and to provide them with a starting place from which to write the final news text.

Our system, adapted from Leppänen et al. (2017a), is based on a pipeline of components with dedicated responsibilities similar to those described by Reiter and Dale (2000) and Reiter (2007). For this work, the relevant part of the architecture is the Document Planner component. This component receives as input two sets of *message* data structures, an example of which is shown in Table 1.[1] The messages are extracted automatically from tables of statistical data obtained from Eurostat.

The *core set* contains messages that are known to be highly relevant to the generation task. Unlike the "summary spine' of Hallett et al. (2006), the set is unlinked and unordered, and not all members of the set are guaranteed to be included in the document plan. The *expanded set*, contains messages that *can* be, but are not guaranteed to be, relevant for the document. Expressed using the terminology from Section 2, we assume that only messages in the core set can be nuclei, while messages from either set can be satellites.

These core and expanded sets are determined automatically from user input. When requesting a new text, the user of the system must define a dataset the text is to be generated from, for example the consumer price data available from Eurostat. This dataset is then divided into the core set and the expanded set by the user when they select what country the generated text should focus on. For example, if the user were to select that the text should discuss French consumer prices, the core set would contain all data from the consumer price dataset that pertains directly to France, while the rest of the consumer price dataset (including data pertaining to the UK, Finland, Croatia, etc.) would be set as the expanded set.

We estimate each message's 'newsworthiness' using the Interquartile Range based method described by Leppänen et al. (2017b) with the values scaled to have mean 0 and standard deviation 1 for the purposes of this computation. The resulting value is conceptually similar to 'importance' of Gatt et al. (2009) and 'risk' of Banaee et al. (2013). The IQR based method compares each data point in turn to a larger distribution, giving it higher scores the further it is from the area between the first and the third quartile of the larger distribution. Values between the quartiles are given a minimal, uniform, score that is dependent on the shape of the distribution. In other words, higher IQR values indicate that the value is more of an outlier compared to the rest of related data in the dataset. As such, it captures a degree of 'unexpectedness', which is an important aspect of newsworthiness (Galtung and Ruge, 1965).

We do not use the domain-specific parts of the method described by Leppänen et al. (2017b). That is, we make no value judgement of whether messages pertaining to French consumer prices are more newsworthy than messages pertaining to Croatian consumer prices, nor do we make judgements of whether changes in the price of education are more or less newsworthy than changes in the price of alcohol and tobacco. However, we do weight the scores so that messages with the `timestamp` field being closer to present receive higher weights, as recency is an important aspect of newsworthiness. While we have described our method for computing the `newsworthiness` value in some detail, we emphasize that for the rest of this article we only assume that the `newsworthiness` values are non-negative and that higher values indicate higher newsworthiness.

More crucially for the method described be-

[1] The concrete implementation details are somewhat more complex. We omit details irrelevant for this work.

low, we specify that the value_type fields (which describe how the messages' values are to be interpreted) contain members of a hierarchical taxonomy of data types represented as colon-separated hierarchies of labels. For example, the value_type field value health:cost:hc2:mio_eur would indicate that the number in the value field is the amount of money (cost), measured in millions of euros (mio_eur), spent by some nation (as defined by the location and location_type fields) on rehabilitative care (hc2) in some time period (as defined by the timestamp and timestamp_type fields) and that this is part of the larger health care topic (health). In our case, these labels are automatically established from the headers of the input data tables.

The goal of document structuring is to produce a three-level tree-structure with ordered children. The root node corresponds to the document as a whole and the mid-level structures correspond to paragraphs. The leaves are the messages selected for inclusion in the document. While the messages have not yet, at this stage, been associated with any linguistic structures, they can be conceptualized as being phrases or very short sentences. We are thus concurrently determining both the content and the structure the document.

We emphasize that our applications are employed in domains where they must be able to select some 10-30 messages from a pool of potential messages numbering in 100,000s. Given infinite computational resources, it would be preferential to construct all possible document plans and then score them in some fashion. This, however, is infeasible given the size of the search space. Previously, other authors have employed, for example, stochastic searches with significantly smaller search spaces (Mellish et al., 1998). Indeed, some kind of a beam search approach could be very useful in smartly searching a subset of the search space. However, we have thus far been unable to identify a document-level metric that adequately balances the 'total amount of newsworthiness' in a text with the length of the text, a requirement for beam search.

4 Research Objective

Based on the above considerations, our main goal is to identify a widely applicable method for content selection and document planning that matches the following requirements:

REQ1: The method needs to be highly performant

REQ2: The method should not be dependent on domain knowledge

REQ3: The document should have a theme

REQ4: The document should have multiple paragraphs but not be excessively long

REQ5: The paragraphs should have distinct themes related to the document theme

REQ6: The paragraph themes should be newsworthy in their own right

REQ7: The paragraphs should not be excessively long or short

REQ8: All messages should relate to the paragraph theme

REQ9: All messages should be newsworthy

REQ10: Within each paragraph, the messages should be presented in an order that produces a coherent narrative

Again, we emphasize that our goal is not to identify a method that is optimal for any specific scenario, but rather to determine a baseline method that is *adequate* for a broad spectrum of applications and sub-domains.

5 A Baseline Approach to Document Planning

Optimally, we would wish to produce some sort of a *globally optimal* document plan. However, as discussed above, this would entail significant computational costs and require a scoring function applicable to the document as a whole. As such, we propose a method for producing document plans in a greedy, linear, and iterative fashion. At every stage, decisions are made considering only a limited local context, thus avoiding the need for a method of determining the global quality of the document plan, thus fulfilling REQ1 ('The method needs to be highly performant').

The document's overall theme, in our use case, is selected by the user who initiates the generation task. In initiating the task, the users selects both a dataset and a focus location. The generation process then derives the *core messages* and *expanded messages* sets (the inputs to the Document Planner, see Section 3) so that both sets discuss the dataset

Field	Description	Example value
`where`	What location the fact relates to	Finland
`where_type`	What the type of the location is	country
`timestamp`	The time (or time range) the fact relates to	2020M05
`timestamp_type`	The type of the timestamp	month
`value`	A (usually) numeric value	0.01
`value_type`	Interpretation of `value`	cphi:hicp2015:cp-hi02:rt01
`newsworthiness`	An estimate of how newsworthy the message is	1

Table 1: An example of a message. The hypothetical message states that in the fifth month of 2020, in Finland, the consumer price index, using the year 2015 as the start of the index, of alcoholic beverages and tobacco changed by 0.01 points with respect to the value of the index during the previous month.

indicated by the user (i.e. messages from other datasets are not generated) and that the core set contains messages pertaining to the user's indicated focus location, while messages pertaining to all other locations are in the expanded set. This fulfills REQ3 ('The document should have a theme'). This step is also independent of the specific subdomain, thus fulfilling REQ2 ('The method should not be dependent on domain knowledge'). This step thus fulfills all the relevant requirements. Next, we'll describe how both the first and subsequent paragraphs can be planned in a way consistent with the requirements defined above.

5.1 Planning the First Paragraph

At the start of the document planning process, we select the most newsworthy message from the *core messages* set to act as the nucleus (n_1) of the first paragraph (p_1). This nucleus establishes the theme of the first paragraph as follows: We inspect the `value_type` field of this first nucleus n_1, and retrieve a prefix `Prefix(`n_1`)`. The prefix is the least amount of colon-separated labels wherein the total amount of prefixes in the core set is greater than the minimal amount of paragraphs a document can have, in our case two. In our case, as a consequence of our label hierarchy, this is always the first three colon-separated units. For the message shown in Table 1, the prefix would thus be `cphi:hicp2015:cp-hi02`, meaning that the first paragraph's theme would be the prices of alcoholic beverages and tobacco. This fulfills REQ5, 'the paragraphs should have distinct themes related to the document theme' for the first paragraph.

Next, the first paragraph is completed with satellites from the union of the *core messages* and the *expanded messages* sets. These satellites are initially filtered so that only messages that have the same prefix as the nucleus n_i are considered in paragraph p_i to fulfill REQ8 ('All messages should relate to the paragraph theme'). The satellites are then selected in a linear, greedy, and iterative manner to fulfill REQ1.

For selecting the k'th satellite to a partially constructed paragraph already containing $k-1$ satellites and one nucleus, we consider both the newsworthiness of the available messages (REQ9), as well as how well they would fit the already constructed segment (REQ8). Observing only the newsworthiness would produce a highly incoherent narrative, whereas focusing only on the narrative risks leaving out highly important information.

Following the reasoning in Section 2, we assume that two subsequent messages are more likely to form a good narrative if they are similar. As such, we need a method for weighing the message's newsworthiness by the similarity of the message to the last message of the under-construction paragraph, thus balancing the requirements of REQ8 and REQ9. In terms of the message objects described in Table 1, it seems to us that the intuitive aspects of similarity are related to the degree of similarity within the 'meta' fields such as `timestamp`, `location` and `value_type`.

For the `timestamp` and `location` fields, we can state that two messages that have identical values in the fields are more similar that two messages that are otherwise the same but have distinct values for said fields. We call this the *contextual* similarity of the messages, and the fields the *contextual fields* (F_c), as these fields provide us access to the larger context in which the `value` and `value_type` fields can be interpreted. Contextual similarity captures the notion that it is likely better to follow a fact about French healthcare spending in 2020 with another piece of information about France in 2020,

rather than about Austria in 1990.

In more precise terms, we propose the following weighing scheme for contextual similarity: The similarity $sim_c(A, B)$ of two messaged A and B is the product of weights $w_f > 1$ for each field f among the contextual fields F_c, where both A and B have the same value for the field:

$$sim_c(A, B) = \prod_{\{f \in F_c | A.f = B.f\}} w_f \quad (1)$$

This value strictly increases as more fields are shared between A and B. We explicitly define the similarity to be zero if there are no fields f where A and B share a value. If w_f is a uniform value for all fields f, this scheme is completely domain-agnostic. Setting different weights w_f for each field $f \in F_c$ allows for encoding some domain knowledge about which fields are the most important for the text, thus providing a method for producing more tailored texts at the cost of slightly violating REQ2. In our case study, we set $w_{timestamp} = 1.1$ and $w_{location} = 1.5$.

The above consideration of similarity still ignores valuable information available from the `value_type` field, which describes how the value in the `value` field is to be interpreted. Denoting `health:cost:hc2:mio_eur` (the cost of rehabilitative care in millions of euros) by T_1, consider its similarity to T_2 = `health:cost:hc2:eur_hab`, the cost of rehabilitative care as euros per inhabitant, and T_3 = `health:cost:hc41:mio_eur`, the cost of health care related imaging services in millions of euros. Intuitively, T_1 and T_2 are thematically closer than T_1 and T_3. We model this similarity between two facts A and B simply as

$$sim_t(A, B) = \frac{1}{s(A, B)} \quad (2)$$

where $s(A, B)$ is the length – in colon-separated units – of the unshared suffix between A and B's `value_type` fields. That is, $s(T_1, T_2) = 1$ whereas $s(T_1, T_3) = 2$. We specify that $sim_t(\cdot, \cdot)$ is zero for all pairs without any shared prefix.

Our formulation of $sim_t(\cdot, \cdot)$ was influenced by the observation that in our context the messages' `value_type` values have a constant number of colon-separated segments. In cases where the lengths of the `value_type` values differ, an alternative formulation of

$$sim'_t(A, B) = \frac{2p(A, B)}{\ell(A) + \ell(B)} \quad (3)$$

where $\ell(\cdot)$ provides the length of the `value_type` value, and $p(\cdot, \cdot)$ is the length of shared *prefix* between A and B, both measured as colon-separated units, might be preferable if also more complex.

When considering whether the k'th satellite s_i^k of paragraph p_i should be a specific candidate $c \in C$, where C is all so far unused messages, we can combine the similarity metrics with the newsworthiness of c into a general fitness value as follows:

$$\begin{aligned} fit(c, x) = c.newsworthiness \\ \times sim_c(c, x) \\ \times sim_t(c, x) \\ \times set\_penalty(c) \end{aligned}$$

The $set\_penalty(c)$ factor depends on whether the message originates from the *core messages* set, or the *extended messages* set. For messages originating from the core message set, the penalty is 1. For messages originating from the extended messages set, the penalty is $\frac{1}{dist+1}$, where $dist$ is the distance from the previous core message.

The final score describing how good of an addition c would be as the kth satellite of the ith paragraph s_i^k is then obtained by taking the average of fitnesses of c in relation to both the nucleus n_i and the previous satellite s_i^{k-1} by computing:

$$score(c, n_i, s_i^{k-1}) = \frac{fit(c, n_i) + fit(c, s_i^{k-1})}{2}$$

This maximizes the newsworthiness of the paragraph's contents (fulfilling REQ9, 'all messages should be newsworthy'), while also enforcing relatedness to the theme of the paragraph (fulfilling REQ8, 'all messages should relate to the paragraph theme') by measuring against the nucleus and with the inclusion of the *set_penalty*. By continuously measuring against the previously selected satellite, the procedure also allows for interludes to e.g. discuss highly newsworthy information related to but not strictly about the paragraph's main topic, or 'thematic drift'. It thus fulfills REQ10 ('Within each paragraph, the messages should be presented in an order that produces a coherent narrative') while also paying attention to the pyramid model of news (See Section 2).

Using $score$, the highest scoring candidate $c_{top} = \arg\max_{c \in C} score(c, n_i, s_i^{k-1})$ is then compared to both an absolute threshold t_{abs} and the newsworthiness of the nucleus n_i multiplied by relative threshold value t_{rel}. Provided that the

maximal paragraph length has not been reached, the top candidate message c_{top} is appended to the paragraph p_i as the k'th satellite s_i^k in the document plan provided that either $score(c_{top}, n_i, s_i^{k-1}) \geq t_{abs}$ or $score(c_{top}, n_i, s_i^{k-1}) \geq t_{rel} \times n_i.newsworthiness$.

These thresholds ensure that the paragraph does not stray into minutiae, whether considered in absolute terms or in relation to the nucleus of the paragraph. In cases where the minimum paragraph length has not been reached, the thresholds are ignored and the top candidate is always appended. This accounts for REQ7 ('The paragraphs should not be excessively long or short').

The above considerations take into account several free parameters, namely the maximal and minimal paragraph lengths as well as the threshold values t_{rel} and t_{abs}. In our case study, we selected the minimal and maximal paragraph lengths as 2 and 5 messages empirically by trialing out various values and observing the resulting texts. These should, naturally, be based on the genre of text and the target audience. For the threshold values we selected 0.2 and 0.5, respectively, using the same method as with the paragraph lengths above. Both the thresholds and the minimal and maximal paragraph lengths should be viewed as (manually) tuneable hyperparameters.

5.2 Planning Subsequent Paragraphs

We then proceed to generate further paragraphs in a manner highly similar to that used when planning the first paragraph. The only distinction is that, when selecting the nucleus n_i for a subsequent paragraph p_i, we obtain the message from the *core messages* set with a highest newsworthiness value that has a prefix (theme) not yet discussed among the previously planned paragraphs p_1 - p_{i-1}:

$$n_i = \arg\max_{c \in C} c.newsworthiness \tag{4}$$

where

$$C = \Big\{ c \in CoreMessages | \texttt{Prefix}(c) \notin \{\texttt{Prefix}(n_k) | k \in [1..i-1]\} \Big\} \tag{5}$$

This ensures that the different paragraphs are highly newsworthy, thus fulfilling REQ6, while also fulfilling REQ5 for having distinct themes for the different paragraphs.

As when constructing the subsequent paragraphs, the total length of the document also needs to be considered. To fulfill REQ4 ('The document should have multiple paragraphs but not be excessively long'), we employ a variation of the method described in the previous section for ending individual paragraphs. A maximal length (in our case, 3 paragraphs) ensures that the document is not allowed to grow beyond reason, whereas a minimal length (for us, 2 paragraphs) ensures that the document is not unreasonably short. After the minimal length has been reached (but not yet the maximal length), a new paragraph is only started if the nucleus of the potential paragraph has a newsworthiness value that is at least 30 % of the newsworthiness value of the first nucleus of the document. This, as with the satellites, ensures that the the document does not stray into minutiae, balancing REQs 4 and 6. the maximal and minimal lengths, as well as the 30 % threshold, were determined by manual fine-tuning and should be viewed as tuneable hyperparameters.

6 Evaluation

The method described above was implemented in a larger NLG application producing news alerts for journalists from datasets provided by Eurostat. A variation of the same application was also developed with a simplified document planner. In this simplified planner, the planner always selects the maximally newsworthy available message as the message without any early stopping threshold. Nuclei are selected from the core messages set, while satellites can be from either set. Contrasting our proposed method with this simplified method enables us to evaluate the importance of narrative coherence in the generated texts. The larger application is multilingual, but the evaluation was conducted using English language texts.

Three experts were recruited from the Finnish News Agency STT, a national European news agency, to evaluate documents on the consumer price indices in five different European nations. For all nations, the judges were shown variants produced by both our proposed method and the simplified method. One of the selected countries is the country the news agency is based in, with the assumption that the judges would have high amounts of world knowledge they would be able to use in evaluating these texts. Another variant pair describes a country that is both relatively small and geographically remote (but still within EU), with the assumption that the journalists are unlikely to

Consumer Prices in Estonia

In June 2020, in Estonia, the monthly growth rate of the harmonized consumer price index for the category 'education' was 30.8 points. It was 30.7 percentage points more than the EU average. In July 2020, it was 0.4 percentage points less than the EU average. It was -0.4 points. In May 2020, the yearly growth rate of the harmonized consumer price index for the category 'education' was -20.5 points. It was 21.9 percentage points less than the EU average.

In August 2020, the monthly growth rate of the harmonized consumer price index for the category 'housing, water, electricity, gas and other fuels' was 2.5 points. It was 2.3 percentage points more than the EU average. In North Macedonia, it was 3 percentage points more than the EU average. It was 3.2 points. Estonia had the 3rd highest monthly growth rate of the harmonized consumer price index for the category 'housing, water, electricity, gas and other fuels' across the observed countries. In Sweden, the monthly growth rate of the harmonized consumer price index for the category 'housing, water, electricity, gas and other fuels' was 3.1 points.

Figure 1: Example output regarding Eurostat statistics on consumer prices. The text contains 12 messages, selected from among 207,210 messages available during generation.

have much world knowledge about this country's consumer prices. The three other countries were selected from among those bordering the first country, with the assumption that the journalists would have some, but not much, world knowledge relating to these countries. The final output texts were not inspected prior to selecting the countries.

All of the texts used in the evaluation were generated from a copy of the same underlying Eurostat dataset, entitled 'Harmonised index of consumer prices - monthly data [ei_cphi_m]'[2] downloaded in September 2020. It contains country-level data regarding the harmonized consumer prices indices, and their change over time, for various EU nations starting from January 1996. We preprocess the data by adding monthly rankings (i.e. determine what country had the greatest, the second greatest, etc. value for a specific index category during any specific month) and comparisons to the EU average values.

As the evaluation was focused on document planning and content selection, the larger system was simplified in some respects, e.g., to not conduct complex aggregation. This was done to minimize the effect of later stages of the generation process on the evaluation. As a result, the language in the evaluated documents was relatively stilted, as exemplified by Figure 1. The only manual alteration was the addition of headings to indicate the texts' intended themes.

The judges did not receive any direct compensation but their employer, the news agency, is a member of the EU-wide EMBEDDIA research project within which parts of this work was conducted. The evaluations were conducted online. The judges were first provided with some basic information on the type of documents they were to read (i.e. that the texts are intended to be news alerts for journalists, rather than publication ready news texts), the length of the task, etc. All instructions were in the judges' native language, in this case Finnish. The judges were not told which texts were produced by which variants nor how many variants were being tested. Following this, the judges were shown the documents one by one. For each document, the judges were asked to indicate their agreement with the following statements (translated from Finnish):

Q1: The text matches the heading

Q2: The text is coherent

Q3: The text lacks some pertinent information

Q4: The text contains unnecessary information

Q5: The text has a suitable length

For Q1–Q4, the judges indicated their agreement on a 7-point Likert scale ranging from 1 ('completely disagree') to 7 ('completely agree'). For Q5, the answers were provided on 5-point scale ranging from 1 ('clearly too short') to 3 ('length is suitable') to 5 ('clearly too long'). In addition, the judges were able to provide textual feedback for each individual text, as well as for the evaluation task as a whole. The judges' answers to Q1 – Q5, are aggregated in Table 2.

The results indicate that the proposed method statistically significantly increases the document's coherence (Q2, mean 4.33 vs. 1.60, median 5 vs 2), the matching of the document's content to the document's theme (Q1, mean 4.40 vs. 1.80, median 5 vs 2), and produces documents of more suitable length (Q5, mean 2.93 vs. 4.07, median 3 vs 4, with 3 being best). The proposed method also seems

[2]Available for download and browsing from `http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ei_cphi_m`

	Our method			Baseline			
Statement	Median	Mean	SD.	Median	Mean	SD.	p_{MWU}
Q1 (1–7, ↑)	5	4.40	1.64	2	1.80	0.41	< 0.001*
Q2 (1–7, ↑)	5	4.33	1.76	2	1.60	0.51	< 0.001*
Q3 (1–7, ↓)	4	4.47	1.81	6	5.80	1.42	0.049
Q4 (1–7, ↓)	5	5.13	1.55	6	6.33	0.62	0.024
Q5 (1–5, 3 best)	3	2.93	0.59	4	4.07	0.70	< 0.001*

Table 2: Results obtained during the evaluation. Parentheses indicate answer ranges and whether the higher (↑), lower (↓) or middle values are to be interpreted as the best. The p_{MWU} column contains the (uncorrected) p-value of a two-sided Mann-Whitney U test. An asterisk indicates the p-value is statistically significant also after applying a Bonferroni correction to account for multiple tests.

to result in less unnecessary information being included in the document (Q4, mean 5.13 vs 6.33, median 5 vs 6), and in the text missing less necessary information (Q3, mean 4.47 vs 5.80, median 4 vs 6), but these effects are not statistically significant after correcting for multiple comparisons with the Bonferroni correction. We hypothesize this difference would become significant in a larger-scale evaluation.

The free-form textual feedback provided by the judges, as expected, indicates that the texts could be further improved. For example, in the case of the text shown in Figure 1, the judges called for a sentence explicitly noting that North Macedonia had the highest monthly growth rate. In addition, they noted it might be better to produce distinct, even shorter, texts as 'news alerts' while reserving the evaluated texts for use as a starting point when the journalist starts writing.

7 Conclusions

In this work, we have identified a need for, and proposed, a widely applicable baseline document planning method for generating journalistic texts from statistical datasets. Our method is based on observations on the similarities between the orbital theory of news structure (White, 1997) and Rhetorical Structure Theory (Mann and Thompson, 1988). While our proposed method is likely to fall short of the performance of subdomain-specific planning methods, results indicate that it achieves adequate performance while fulfilling a set of requirements identified based on the larger application domain of news generation.

Acknowledgements

This work is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media), and grant agreement No 770299, project NewsEye (A Digital Investigator for Historical Newspapers).

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512.

Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. 2013. Towards NLG for physiological data monitoring with body area networks. In *14th European Workshop on Natural Language Generation, Sofia, Bulgaria, August 8-9, 2013*, pages 193–197.

David Caswell and Konstantin Dörr. 2018. Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism practice*, 12(4):477–496.

Seniz Demir, Sandra Carberry, and Kathleen F. McCoy. 2010. A discourse-aware graph-based content-selection framework. In *Proceedings of the 6th International Natural Language Generation Conference*.

Laurence Dierickx. 2019. Why news automation fails. In *Computation+ Journalism Symposium, Miami, FL*.

Konstantin Nicholas Dörr. 2015. Mapping the field of algorithmic journalism. *Digital journalism*.

Longxu Dou, Guanghui Qin, Jinpeng Wang, Jin-Ge Yao, and Chin-Yew Lin. 2018. Data2text studio: Automated text generation from structured data. In

Proc. 2018 Conference on Empirical Methods in Natural Language Processing.

Ondřej Dušek, David M Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 421–426.

Alexander Fanta. 2017. Putting Europe's robots on the map: automated journalism in news agencies. *Reuters Institute Fellowship Paper*, pages 2017–09.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. *arXiv preprint arXiv:1908.09022*.

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):64–90.

Albert Gatt, Francois Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *Ai Communications*, 22(3):153–186.

Dimitra Gkatzia. 2016. Content selection in data-to-text systems: A survey. *arXiv preprint*. Available at `https://arxiv.org/abs/1610.08375`.

Andreas Graefe. 2016. Guide to automated journalism.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Catalina Hallett, Richard Power, and Donia Scott. 2006. Summarisation and visualisation of e-health data repositories. In *UK E-Science All-Hands Meeting*.

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

Ioannis Konstas and Mirella Lapata. 2013. Inducing document plans for concept-to-text generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1503–1514.

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017a. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197.

Leo Leppänen, Myriam Munezero, Stefanie Sirén-Heikel, Mark Granroth-Wilding, and Hannu Toivonen. 2017b. Finding and expressing news from structured data. In *Proceedings of the 21st International Academic Mindtrek Conference*, pages 174–183. ACM.

Liunian Li and Xiaojun Wan. 2018. Point precisely: Towards ensuring the precision of data in generated texts using delayed copy mechanism. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1044–1055, Santa Fe, New Mexico, USA. ACL.

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

Veronika MacKová and Jakub Sido. 2020. The robotic reporter in the Czech News Agency: Automated journalism and augmentation in the newsroom. *Communication Today*, 11(1):36–53.

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

Chris Mellish, Alistair Knott, Jon Oberlander, and Mick O'Donnell. 1998. Experiments using stochastic search for text planning. In *Natural Language Generation*.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proc. 33rd AAAI Conference on Artificial Intelligence*.

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

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 97–104. Association for Computational Linguistics.

Ehud Reiter. 2018. Hallucination in neural NLG. `https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/`. Accessed: 2020-03-02.

Ehud Reiter. 2019. ML is used more if it does not limit control. `https://ehudreiter.com/2019/08/15/ml-limits-control/`. Accessed: 2020-07-25.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Studies in Natural Language Processing. Cambridge University Press.

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

Somayajulu G Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003. Generating English summaries of time series data using the Gricean maxims. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 187–196.

Olli Sulopuisto. 2018. Uutisia kortti kerrallaan. *Suomen Lehdistö*. `https://suomenlehdisto.fi/uutisia-kortti-kerrallaan/`.

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

Peter White. 1997. Death, disruption and the moral order: the narrative impulse in mass-media 'hard news' reporting. *Genres and institutions: Social processes in the workplace and school*, 101:133.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*.

Yleisradio. 2018. Avoin voitto. `https://github.com/Yleisradio/avoin-voitto`.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. ACL.