Placing mined clues on causality at the heart of narrative planning

Pierre-Luc Vaudry and Guy Lapalme
RALI-DIRO – Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal, Québec, Canada, H3C 3J8.
email: {vaudrypl,lapalme}@iro.umontreal.ca

January 9, 2018

Abstract

Narrative data-to-text systems seem to acknowledge causal relations as important. However, they play only a secondary role in their document planners and their identification relies mostly on domain knowledge. This paper proposes an assisted temporal data interpretation model by narrative generation in which narratives are structured with the help of a mix of automatically mined and manually defined association rules. The expressed associations act as clues that suggest causal hypotheses to the reader, who can thus construct more easily a causal representation of the events. Sequential association rules are selected based on the criteria of confidence and statistical significance as measured in training data. World and domain knowledge association rules are based on the similarity of some aspect of a pair of events or on causal patterns difficult to detect statistically. To report about a specific period, pairs of events for which an association rule applies are linked to form an associative network. The model selects the most unusual facts assuming that an event implied by another one with a relatively high probability may be left implicit in the text. The structure of the narrative, called the connecting associative thread, forms a spanning tree over the selected associative sub-network, which is then segmented into paragraphs and sentences. The microplanning step assembles event descriptions with discourse markers and appropriate anaphoric expressions. Human evaluation shows that the texts are understandable and the lexical choices adequate.

1 Introduction

Whether they are standard on now ordinary devices such as mobile phones or are specialized for healthcare or surveillance purposes, for example, sensors of all kinds record more of our lives every day. Industrial equipment monitoring also generates considerable amounts of data. This adds to the data accumulating on computers around the globe on commercial

and financial transactions, web traffic, and click through statistics. These examples have in common large volumes of frequently updated non-textual event data. Much of these temporal data are heterogeneous and involves actions performed or experienced by individuals or groups. To present this information and allow the concerned persons to quickly understand the situation and make better decisions, computers must be able to discover links between events and express them properly.

An attractive way of presenting real-life temporal data to help in its interpretation is an automatically generated narrative. To approach the writing ability of a human expert, one should probably take into account that narrative comprehension involves the construction of a causal network by the reader [32, 33]. This paper exploits this and proposes an assisted temporal data interpretation model by narrative generation.

This model structures a narrative by relying on a mix of automatically mined and manually defined association rules. The first are collected using sequential association data mining techniques. The second come from formalized world and domain knowledge. The extracted associations do not correspond directly to causal relations. Rather they function as hints for the reader to form causal hypotheses. There is no guarantee that all associations will help in forming a correct causal network. Some associations may be non-causal and constitute red herrings that the reader has to ignore. The generated text's communicative goal is to help the reader assimilate the facts necessary to construct a causal representation of the events. The connecting associative thread that structures the text guides the reader from its beginning to its end.

The proposed model of assisted temporal data interpretation by narrative generation should be applicable in any scenario where:

- events happen repetitively enough to accumulate statistics;
- a summary of a given period is required on a regular basis to monitor unusual events;
- it would not be worthwhile to ask a human to go through the detailed sequence of events to write the summary themselves.

Growing quantities of data corresponding to the above criteria are available. Mobile devices equipped with ever more numerous and varied sensors occupy an increasingly large place in our lives [17]. For example, researchers currently study their daily use for the monitoring of the health of a person with a chronic disease or a mental disorder [23, 24]. Specialized sensors could also be applied temporarily during a medical emergency to allow better transfer of information in the health-care chain (paramedic, nurse, doctor, etc.) [28]. Even without specialized sensors, our behavior is often recorded. Video monitoring, whether for reasons of health, safety or traffic monitoring, produces a lot of data that need to be analyzed and synthesized in order to detect problematic behaviors [18]. All commercial and financial transactions recorded daily can provide important information on consumer behavior and the economy, which is valuable for business, government and society in general. On the web, the millions of clicks recorded have the potential to improve the user experience by better identifying their needs. Finally, the various logs recorded by computer servers need

to be exploited to better manage traffic and network security. These are all situations in which a textual digest of events would be useful and this paper shows a way to achieve this goal.

The paper is organized as follows. After reviewing previous work, a section is dedicated to presenting the assisted temporal data interpretation model. Next its components are detailed, beginning in Section 4 with those that deal with extracting an associative network from the data. Section 5 complements the model's detailed description with the presentation of the components that exploit this associative network to generate a narrative. Finally, the evaluation of the model is discussed in Section 6.

2 Previous work

A narrative is a text presenting with a certain point of view a series of logically and chronologically related events caused or experienced by actors [3]. Some have pointed at causal relations as a means of improving the narrative aspect of temporal data-to-text generation [13, 6].

The concepts of causal network and causal chain have been used to explain the process of narrative comprehension in humans, see Trabasso *et al.* [32, 33]. Those causal networks are essentially composed of physical and mental events and states (of which goals and actions) connected by causal relations. Restrictions apply on which types of causal relation can connect which types of event or state. The causal chain comprises the events that are on a path traversing the causal network from the introduction of the protagonists and setting to either goal attainment or the consequences of failure. Being on a causal chain and having more causal connections have both been found to increase chances of an event being recalled, included in a summary or judged important by the reader. It is noteworthy that causal networks have been applied to the automatic creation of fictional stories [29, 31].

Data-to-text systems taking temporal data as input, for their part, tend to use handcrafted expert rules to identify causal relations [11, 26, 13, 36, 4, 25]. Another way of encoding domain knowledge on causal relations is the task model of Baez Miranda *et al.* [1, 2]. By contrast, in the meteorological domain, multiple regression and machine learning are also employed to identify causal relations [36].

Causal relations are used in different ways in the process of generating a narrative from data. [11, 26, 13] plan the document by selecting key events and adding some related events. Causal relations are just one way events can be related. Wanner *et al.* [36] and Bouayad-Agha *et al.* [4] use lexical and semantic rules to vary how semantic relations, including causal relations, are expressed in the text. Ponnamperuma *et al.* [25] plan the text by specifying a schema and Baez Miranda *et al.* [2] select the content based on their task model.

Generally, in narrative data-to-text, causal relations are used and acknowledged as important, but they do not play a central role in planning the text. Furthermore, most systems have a domain-specific planning algorithm or a generic planning model that requires a lot of effort to be instantiated for a particular domain. The next section presents a narrative data-to-text model that addresses those issues.

3 Assisted Temporal Data Interpretation Model

The model presented in this paper is a data-driven model for generating narratives for assisting human interpretation of temporal data. It features a discourse structure aiming at leading the reader from the beginning to the end of the narrative. It assumes that the process of narrative comprehension involves the mental construction of a causal network by the reader.

The model includes the extraction of sequential association rules using data-mining techniques [12]. This means that patterns where it does not seem plausible that luck can explain an event type frequently following another event type are gathered from training data. Those sequential association rules are used for event selection and document structuring. The use of data mining techniques to capture information potentially useful for causal interpretation allows both to rely less on domain knowledge and to better adapt to the characteristics of a specific dataset. Additionally, no data-text pairs are needed, as would have been the case with supervised learning-based generation.

By association, we mean a connection between events or states without specifying the nature of the underlying relation. For example, an association can be based on a frequent sequence or a strict similarity. For the purpose of narrative comprehension, we assume that interesting associations are those that can help the reader formulate causal hypotheses. However, a particular association identified in the data is not guaranteed to be causal in nature. That is why the common sense and expertise of the reader are needed to assemble an adequate causal representation.

In contrast with work on reader inference in fictional narrative generation such as Niehaus and Young [21], in our model the computer generating the text does not have access to causal relations or even to all relevant real-world events. Relying on predictions of what the reader will infer when reading the text is not possible, because the computer is missing world and domain knowledge that only the human reader can possess. Instead, the computer finds associations in the data and presents them in a form that the human reader can interpret as a narrative.

Note that although this is not a model for creating fictional narratives, its function is to suggest new associations between previously disassociated events. In this sense and to the extent that it accomplishes this, it can be considered to produce original, creative text [14].

Figure 1 gives an overview of our model to help understand the different sources of information that are taken into account. This is not a flowchart as some steps are not formally specified, but this figure can be used as an outline of the paper. We will refer to its components by using numbers for steps and letters for representation levels. Association rules come from two sources: data mining (1) for sequential association rules (B) from training data (A) and world and domain knowledge (C) formalized as rules (D). The data about a specific period (E) is interpreted (2) using the association rules to create an associative network (F). Then a sub-network containing the most unusual facts (G) is selected (3) using the probabilities of the corresponding sequential association rules (B). The following step of document structuring (4) involves determining the connecting associative thread going from the beginning to the end of the narrative (H). Microplanning (5) produces from this the

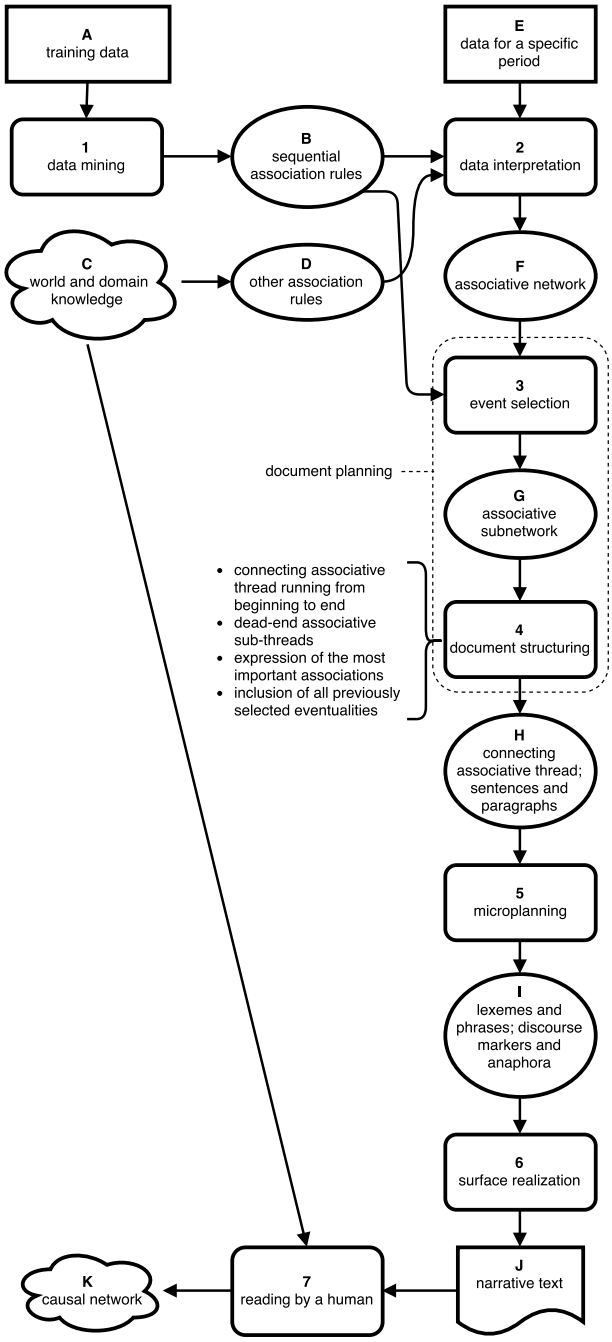


Figure 1: Assisted temporal data interpretation model. Rectangles represent input data; rounded rectangles: computational representations; ellipses: steps; clouds: hypothesized mental representations; rectangle with S-shaped bottom side: a natural language document. For ease of reference, steps are identified by a number and representations by a letter.

lexico-syntactic specification (I). This specification is then realized (6) as a text (J) read by a human (7). The human reader uses his knowledge (C) to reason about the associations expressed in the text. From this they form a mental representation which hypothetically includes a form of causal network (K).

We illustrate our model by generating a narrative from extracted associations and presenting unusual facts from a simple yet realistic dataset, the UCI ADL (Activities of Daily Living) Binary Dataset [22]. This dataset was assembled to train activity classifiers that take as input raw sensor data. This dataset includes the ADLs of two users (A and B) in their own homes. The data was recorded for 14 and 21 consecutive days, respectively. Binary sensor events and the corresponding activity labels are given. We used only the latter in our experiments. For each sensor event or activity, the start and end time are given. There is no overlap between sensor events and between activities (there was only one person per house).

The ADL label set is: *Leaving*, *Toileting*, *Showering*, *Sleeping*, *Breakfast*, *Lunch*, *Dinner*, *Snack*, *Spare_Time/TV*, *Grooming*. The ADL sequence for user A comprises 248 activities (average of 18 activities per day) and that for user B, 493 activities (average of 21 activities per day). To illustrate the various representation levels of the model, an example based on this dataset is provided throughout the remaining of this paper. The input for this running example consists in the data for user B as training data and the portion covering the day of November 24, 2012 as the data to describe. The two first columns of Figure 2 list the start time and the ADL labels of the activities.

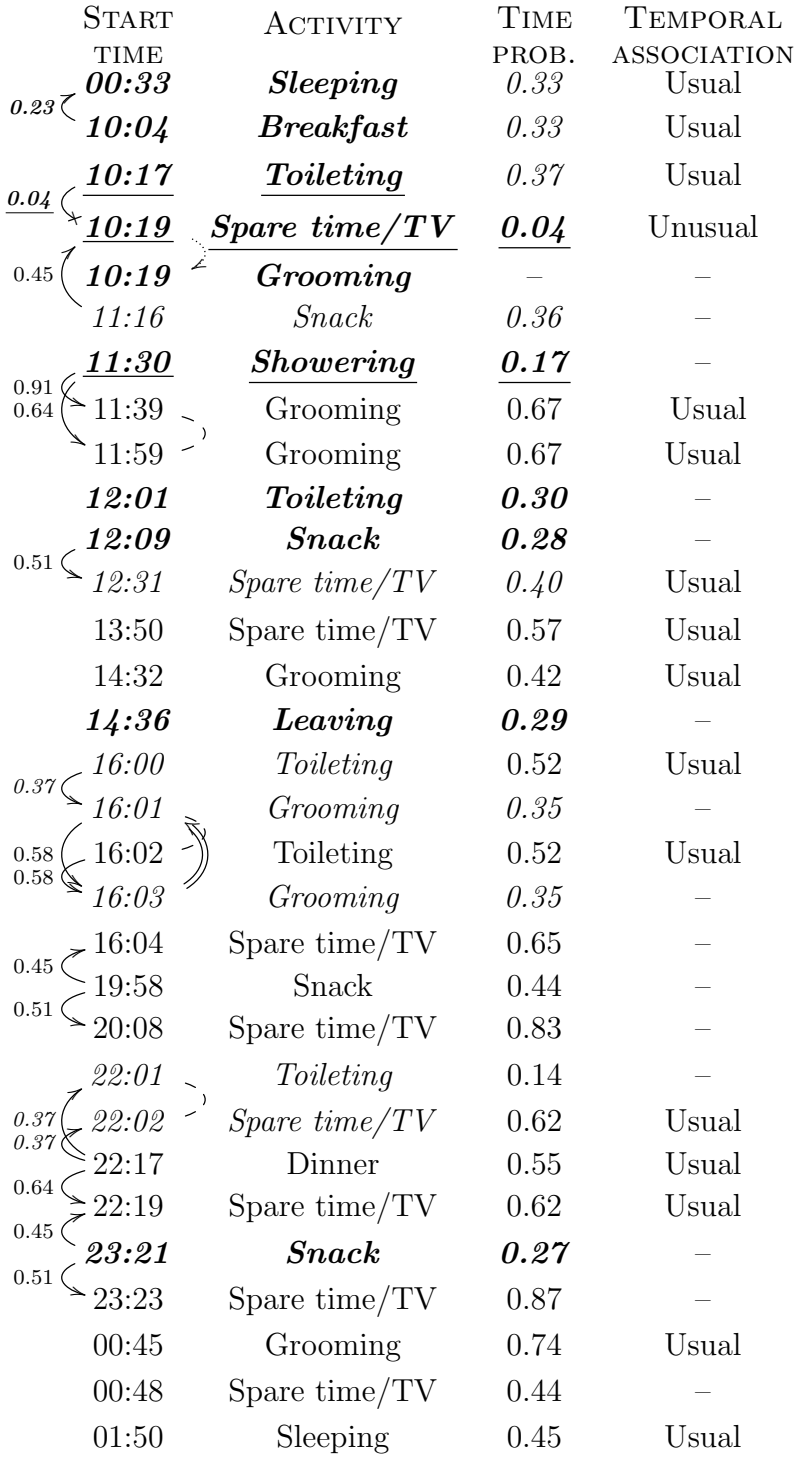
In a real application and in an ideal experiment, the training data would not include the data about the day to cover and subsequent days. Predictive relationships about the user's routine would be extracted from past data in the form of sequential association rules. This knowledge would then be used to evaluate the likelihood that events in the current period follow the same routine. However, considering the small size of the dataset, including all the user's data in the training data helped mitigate data sparsity problems. That way, texts could be generated for any day of the dataset, instead of only the last few days for each user.

4 Associative Network Extraction

This section explains how an associative network is extracted from temporal data according to our model.

4.1 Sequential Association Rule Mining

For finding significant sequential association rules in the ADL data (step 1 on Figure 1), we used the data mining techniques presented by Hamalainen and Nykanen [12], as explained below. This approach was selected because it has been successfully applied for the construction of a causal network from a surveillance video [16]. The video was first segmented spatially and temporally using only pixel information to form the nodes of the network. The causal network was then presented as a visual (non-textual) summary of the video.



Start time	Activity	Time prob.	Temporal association
00:33	***Sleeping***	*0.33*	Usual
10:04	***Breakfast***	*0.33*	Usual
10:17	***Toileting***	*0.37*	Usual
10:19	***Spare time/TV***	***0.04***	Unusual
10:19	***Grooming***	–	–
11:16	*Snack*	*0.36*	–
11:30	***Showering***	***0.17***	–
11:39	Grooming	0.67	Usual
11:59	Grooming	0.67	Usual
12:01	***Toileting***	***0.30***	–
12:09	***Snack***	***0.28***	–
12:31	*Spare time/TV*	*0.40*	Usual
13:50	Spare time/TV	0.57	Usual
14:32	Grooming	0.42	Usual
14:36	***Leaving***	***0.29***	–
16:00	*Toileting*	0.52	Usual
16:01	*Grooming*	*0.35*	–
16:02	Toileting	0.52	Usual
16:03	*Grooming*	*0.35*	–
16:04	Spare time/TV	0.65	–
19:58	Snack	0.44	–
20:08	Spare time/TV	0.83	–
22:01	*Toileting*	0.14	–
22:02	*Spare time/TV*	0.62	Usual
22:17	Dinner	0.55	Usual
22:19	Spare time/TV	0.62	Usual
23:21	***Snack***	***0.27***	–
23:23	Spare time/TV	0.87	–
00:45	Grooming	0.74	Usual
00:48	Spare time/TV	0.44	–
01:50	Sleeping	0.45	Usual

Figure 2: Associative network for user B on November 24, 2012. Sequential associations are on the left. The X-headed arrow represents an unexpected association. On the right are *Instead* (dotted), *Conjunction* (dashed), and *Repetition* (double). Under *Time prob.* is the confidence of the corresponding temporal association rule candidate. If the latter was selected as an expected or unexpected rule, it is marked as *Usual* or *Unusual*, respectively. *Same category* associations are not shown. Selected events are shown for maximum probability 0.2 (underlined), 0.3 (bold), and 0.4 (italics).

In our experiments we considered a limited number of simple types of association rules in the ADL data. To select them we assumed that **temporal proximity was an indicator of potential causality**. This is linked with the covariation principle of attribution theory [15]. According to this social psychology theory, people tend to attribute an effect to a possible cause if they covary. This implies temporal proximity, because there must be observed instances where both event types are present and where both are absent.

Temporal proximity is far from being a guarantee of causality, but it is simple to apply. An association based on that assumption could reflect a direct or indirect causal relation, imply a common cause between events, or be completely unrelated to causality. Whatever the case, it is relevant to express associations of that kind in the generated text, because it gives the reader clues to help assemble a causal representation, even if some are red herrings. Associations extracted from sensor data can only be imperfect, because sensor data contain only a fraction of the relevant information. However, what counts in the end is the causal representation the human reader reconstructs in his mind with the help of other sources of information, not the associations the machine suggested.

A sequential association rule $X_{p-1} \rightarrow Y_p$ means that an event of type Y (a categorical variable) tends to follow sequentially an event of type X, p designating a position in the sequence of events. We also use the notation $X_{p-1} \leftarrow Y_p$ to mean that an event of type X tends to precede an event of type Y. The arrow represents a probabilistic implication. That is, an implication that holds with a certain probability [9]. The direction of this probabilistic implication is important later on in the NLG pipeline. In event selection, an event implied with a certain minimum probability by another event will not be selected to be part of the explicit content of the generated text (see Section 5.1). This assumes that the sequential association rules are a good approximation of the knowledge of a human reader about the situation. If this is correct, the reader could guess the occurrence of the implied event from the mentioned event. However, this should only happen if the implied event is necessary for understanding the text [21]. There is no clear example of the latter in the running example.

Temporal association rules are a special type of sequential association rule. They include a categorical temporal variable such as the hour of the day, day of the week or month. For the purpose of finding association rules between values of such a variable and real event types, a dummy event is created for each time step to indicate the value of the temporal variable.

The types of sequential association rule considered are shown in Table 1. In the following, A and H are categorical variables and stand respectively for activity and hour of the day (hours 0-23, not considering minutes). $A_{i,p}$ stands for a particular type of activity i at position p in the event sequence. Association rule type 1 evaluates the influence of the last activity on the choice of the current activity and vice versa. Type 2 does the same for the penultimate activity and type 3 for the last two activities. Type 4 takes into account the influence of the current hour of the day on the choice of activity. Lastly, type 5 verifies the presence of an association between the combination of the current hour and last activity and the current activity. Each rule is accompanied by an example with the first *Toileting* activity of Figure 2. Rules 1, 2, and 3 are justified by the previously mentioned hypothesis

Type	Association rule	Example candidate
1	$A_{i,p-1} \underset{(\leftarrow)}{\rightarrow} A_{j,p}$	$A_{Breakfast,p-1} \underset{(\leftarrow)}{\rightarrow} A_{Toileting,p}$
2	$A_{i,p-2} \underset{(\leftarrow)}{\rightarrow} A_{j,p}$	$A_{Sleeping,p-2} \underset{(\leftarrow)}{\rightarrow} A_{Toileting,p}$
3	$A_{i,p-2} \wedge A_{j,p-1} \underset{(\leftarrow)}{\rightarrow} A_{k,p}$	$A_{Sleeping,p-2} \wedge A_{Breakfast,p-1} \underset{(\leftarrow)}{\rightarrow} A_{Toileting,p}$
4	$H_{i,p} \underset{(\leftarrow)}{\rightarrow} A_{j,p}$	$H_{10,p} \underset{(\leftarrow)}{\rightarrow} A_{Toileting,p}$
5	$A_{i,p-1} \wedge H_{j,p} \underset{(\leftarrow)}{\rightarrow} A_{k,p}$	$A_{Breakfast,p-1} \wedge H_{10,p} \underset{(\leftarrow)}{\rightarrow} A_{Toileting,p}$

Table 1: Sequential association rule types and candidate examples. The arrows in parentheses indicate that the direction of the probabilistic implication is still undetermined at this point.

that temporal proximity is an indicator of a potential relation of causal nature. The last activity is very close to the current one and the second last is generally also not so far. Rules 4 and 5 are justified by the cyclic nature of temporal phenomena. In the case of reporting a day of ADLs, it seems reasonable to assume that a person will tend to do similar things in the same part of each day. In other words, people tend to follow more or less regularly a daily routine. The choice of the hour as the unit, rather than a finer grain unit such as the minute, is dictated by the relatively small size of the dataset and the need to have enough occurrences for each value of the variable for the statistics to be reliable. A coarser unit such as morning/afternoon/evening was not chosen because it risked being too imprecise. This is because each individual gets up, eats, goes out and goes to bed at different hours of the day, thus tending to shift the notions of morning, afternoon and evening.

For selecting significant sequential association rules, three properties were computed for each candidate [12]:

frequency : the probability of encountering an instance of the association rule in the data; it is estimated from counts;

confidence : the conditional probability of encountering an instance of the association rule, given that an instance of the left hand (chronological direction) or the right hand (reverse chronological) of the association rule is encountered;

significance : the probability of obtaining the observed counts if the events on the right side of the rule were actually independent of the events on the left side of the rule. It is measured by computing the p-value according to the binomial distribution.

The chronological direction of each candidate sequential association rule is determined by computing the confidence for the two possible directions (chronological and reverse chronological) and retaining the direction with the highest one. That means that for candidate association AB, the algorithm checked which could be predicted with more confidence: that given the presence of A, B follows it or that given the presence of B, A precedes it. This

enables a better estimation of the unusualness of each fact and thus improves content selection. This is in the case of rules referring to activities. In the case of temporal association rules, the direction of the rule is not specific, because the hour of the day is by definition simultaneous to the current activity. The confidence considered hereafter for each candidate sequential association rule is the one corresponding to the determined direction (the highest one). For example, the candidate rule $A_{Sleeping,p-1} \rightarrow A_{Breakfast,p}$ has $cf = 0.17$ and its reverse chronological counterpart $A_{Sleeping,p-1} \leftarrow A_{Breakfast,p}$ has $cf = 0.23$. Consequently, only the second one is retained and will be used to estimate the probability of this sequence.

Two p-values were computed to establish the significance of each candidate: $p_{expected}$, to indicate positive association rules (significantly high counts), and $p_{unexpected}$, to indicate negative association rules (significantly low counts). By the latter we mean cases in which the presence of the events on the left side of the rule can be used as a predictor of the absence of the events on the right side of the rule. In other words, actual instances of these association rules are unexpected.

Frequency, confidence and significance are formalized in Figure 3.

To compute frequency, confidence and significance, we counted in the data $m(L_i, R_j)$ and $m(L_i)$ for each value i, j for each association candidate $L_i \rightarrow R_j$. Those counts were made using all the data available for a given user.

Next, the association rule candidates are filtered using the following criteria. To get the expected association rules, only candidates $L_i \rightarrow R_j$ for which $cf(L_i \rightarrow R_j) > cf_{min}$ and $p_{expected}(L_i \rightarrow R_j) < 0.05$ were retained. To get the unexpected association rules, only candidates $L_i \rightarrow R_j$ for which $cf(L_i \rightarrow R_j) < cf_{max}$ and $p_{unexpected}(L_i \rightarrow R_j) < 0.05$ were retained. We tried different values of cf_{min} and cf_{max} and settled for $cf_{min} = 0.3$ and $cf_{max} = 0.07$. This seemed reasonable because there were 10 different ADL labels (e.g. *Sleeping*, *Grooming*, etc.) such as the ones shown in Figure 2, which would give an a priori probability of 0.1 for each without any knowledge about the data. This means that associations that have a conditional probability of having their right side happening with a probability around 0.1 given their left side do not give much information. They are thus less relevant.

Candidates also had to be filtered to eliminate redundancy: $L_i^1 \rightarrow R_j$ is considered more general than $L_k^2 \rightarrow R_j$ if and only if the events of L_i^1 are included in the events of L_k^2. For example, the rule $A_{Breakfast,p-1} \rightarrow A_{Toileting,p}$ is more general than $A_{Sleeping,p-2} \wedge A_{Breakfast,p-1} \rightarrow A_{Toileting,p}$. A rule candidate was considered non-redundant only if all more general rule candidates were less significant (had a higher p-value). A more general rule candidate was still kept too if it was significant enough (p-value < 0.05).

For example, among the five rule candidates with *Toileting* given in Table 1, only $H_{10,p} \rightarrow A_{Toileting,p}$ ($cf = 0.365$, $p_{expected} = 0.002$) was selected as an expected association rule and none as an unexpected association rule. An example of a rule candidate that was selected as an unexpected rule is $A_{Toileting,p-1} \wedge H_{10,p} \rightarrow A_{Spare\_Time/TV,p}$ ($cf = 0.044$, $p_{unexpected} = 0.028$). Those numbers come from the counting of all the 21 days of data available for user B.

Rules 1 to 5 of Figure 4 are examples of mined sequential association rules.

Count for value i of variable X: $m(X_i)$
Total count for variable X:

$$n(X) = \sum_i m(X_i)$$

Probability for value i of variable X:

$$P(X_i) = \frac{m(X_i)}{n(X)}$$

Joint count for values i, j of left (L) and right (R) parts of association rule $L_i \rightarrow R_j$:

$$m(L_i, R_j)$$

Total joint count for an association rule of type $L \rightarrow R$:

$$n(L, R) = \sum_{i,j} m(L_i, R_j)$$

Frequency of an association rule $L_i \rightarrow R_j$:

$$fr(L_i \rightarrow R_j) = P(L_i, R_j) = \frac{m(L_i, R_j)}{n(L, R)}$$

Confidence of a chronological association rule $L_i \rightarrow R_j$:

$$cf(L_i \rightarrow R_j) = P(R_j|L_i) = \frac{P(L_i, R_j)}{P(L_i)}$$

Confidence of a reverse chronological association rule $L_i \leftarrow R_j$:

$$cf(L_i \leftarrow R_j) = P(L_i|R_j) = \frac{P(L_i, R_j)}{P(R_j)}$$

Significance (p-value using the binomial distribution) of $L_i \rightarrow R_j$:

$$p(L_i \rightarrow R_j) = \sum_{l=l_{min}}^{l_{max}} \binom{n(L,R)}{l} (P(L_i)P(R_j))^l (1 - P(L_i)P(R_j))^{n(L,R)-l}$$

With $l_{min} = m(L_i, R_j)$ and $l_{max} = m(L_i)$ for an expected association ($p_{expected}$) and $l_{min} = 0$ and $l_{max} = m(L_i, R_j)$ for an unexpected association ($p_{unexpected}$).

Figure 3: Notation and formulas for counts, frequency, confidence, and significance.

Mined sequential association rules:

1. $H_{11,p} \rightarrow A_{Grooming,p}$
   $cf = 0.67, p_{expected} = 0.000005$

2. $A_{Sleeping,p-1} \leftarrow A_{Breakfast,p}$
   $cf = 0.23, p_{expected} = 0.01$

3. $A_{Showering,p-2} \rightarrow A_{Grooming,p}$
   $cf = 0.64, p_{expected} = 0.01$

4. $A_{Grooming,p-2} \wedge A_{Toileting,p-1} \rightarrow A_{Grooming,p}$
   $cf = 0.58, p_{expected} = 0.001$

5. $A_{Toileting,p-1} \wedge H_{10,p} \not\rightarrow A_{Spare\ time/TV,p}$
   $cf = 0.04, p_{unexpected} = 0.03$

World and domain knowledge association rule:

6. $A_{i,p} \xleftrightarrow{Same\ category} A_{j,q} \iff category(i) = category(j)$

Figure 4: Association rule examples. A and H are categorical variables and stand respectively for activity and hour of the day (hours 0-23, not considering minutes). A_i, p stands for a particular type of activity i at position p in the event sequence. cf stands for confidence. $p_{expected}$ and $p_{unexpected}$ are p-values that measure the significance of expected and unexpected association rules, respectively (lower is better).

4.2 World and Domain Knowledge Rules

Causally relevant world and domain knowledge can be formalized as association rules (C and D in Figure 1). Association rules can be based on three kinds of similarities between events: similarity of event types, similarity of event arguments, and similarity of event circumstances. Event arguments are any arguments of the predicate corresponding to an event type. This includes roles such as agent, patient, instrument, etc. Circumstances are characteristics of the context of an event that are not part of its arguments, such as location, weather, surroundings, manner, etc.

Note that those kinds of similarity explicitly cover two dimensions of the five included in the narrative comprehension situation model of Zwaan, Langston, and Graesser [37]: spatiality and protagonist. The other three situation dimensions that they use in modelling textual discontinuities are temporality, causality and intentionality. Continuity in temporality is used in this model as a default association in building the connecting associative thread (see Section 5.2.1). Causality and intentionality are hypothesized to underlie part of the associative network.

A form of similarity of event arguments has been employed in generating narratives by Gervás [6]. In the context of a story with multiple actors, he associates actions having the same actor. In this way the narrative is focussed around the perceptions and actions of one actor at a time. Of course, this is not needed in the case of data featuring only one actor like the dataset used in our running example.

World and domain knowledge is used to define which dimensions of similarity are relevant and how they should be evaluated in a particular application. Similarity-based association rules can be entered individually or come from an existing ontology which could be used to evaluate the similarity of two event types, arguments or circumstances.

The associations derived from similarity-based association rules have the advantage of linking events regardless of their place in the sequence. That means that they can be used to create long-distance links in the text while keeping temporally close events also close in the text. Only a small number of similarity-based association rules are necessary if they are general enough to apply to all event types. In this way a good proportion of event pairs will be associated and the associative network will have more chances to be connected. This will result in a more appropriate discourse structure, with fewer event pairs lacking an explicitly marked discourse relation. See Section 5.2.1 for more details on the construction of the connecting associative thread, the narrative discourse structure proposed in this model.

Rule 6 of Figure 4 is a simple but effective example of a manually entered association rule based on similarity of the event type. It defines a *Same category* association. For the purpose of the ADL example, we arbitrarily grouped the ADL types into categories. *Toileting*, *Grooming*, and *Showering* were placed in the category of personal hygiene activities. *Breakfast*, *Lunch*, *Dinner*, and *Snack* were grouped as eating activities. *Spare_Time/TV*, *Leaving*, and *Sleeping* were kept in separate categories, because they were considered to have significantly different functions and locations from the others and each other.

In the example ADL sequence of Figure 2, this similarity-based association rule creates associations between events that are not necessarily sequentially or temporally close, such

as the 10:04 *Breakfast* and the 12:09 *Snack* or the 16:02 *Toileting* and the 22:01 *Toileting*. Those associations would not have been covered by the sequential association rule type of Table 1. In this way the *Same category* association rule enables long-distance links between ADL activities and increases the proportion of events that are connected with each other. This can help obtain a more appropriate connecting associative thread in document planning.

Manually defined association rules may also be necessary to take into account specific pieces of knowledge difficult to capture with mined sequential association rules. Some causally relevant patterns can be impractical to derive from temporal proximity and relative frequencies in the limited training data. An example of specific causal patterns would be the hand-authored commonsense axioms of Gordon [7]. Those axioms encode some of the world knowledge necessary to interpret the behavior of human-like agents in a simplified context. They include an estimation of the likelihood that the antecedent implies the consequent. A given set of association rules could, for example, present plausible causes for an agent x to chase an agent y. One of these rules could state that if x is playing with y, x may chase y with a likelihood of 0.2. Another rule could say that if x has the goal of making y afraid of x, x may chase y with a likelihood of 0.5.

4.3 Data Interpretation

Data interpretation (step 2 of Figure 1) consists in searching the data for instances where an association rule applies. Sequential associations are derived from rules such as Rules 2 to 5 from Figure 4. They are shown as arrows going from one row to another at the left of Figure 2. The arrow labels indicate the confidence of the corresponding sequential association rule. Temporal associations are derived from rules such as Rules 1 and 5 from Figure 4. They are indicated by the *Time prob.* and *Temporal association* columns in Figure 2. *Usual* means that an expected association was found and *Unusual* indicates an unexpected association. No indication means that time was not considered significantly useful in predicting those occurrences (no association rule). The probability conditional on time (the confidence of the corresponding association rule candidate) is in any case indicated as it will be used for content selection.

When we decided to try mining reverse chronological sequential association rules, our hypothesis was that it could help capture underlying goals having a later manifestation. That means that for an event sequence AB, if it is easier to predict a preceding A knowing B than a following B knowing A, according to this hypothesis, it could be because the goal of A was to make B happen. For example, consider the reverse chronological associations *Dinner 22:17* to *Toileting 22:01* and *Spare time/TV 22:02* shown in Figure 2. One could imagine that the latter two activities were accomplished in preparation for *Dinner 22:17*. It would seem plausible that the user would have been to the toilet and waited in the living room until it was time for dinner. This is because it can be assumed that the *Dinner* activity occurrences are predictable enough that one could expect some preparation for it. However, the underlying goal hypothesis does not seem to hold in all cases. Again by looking at Figure 2, a reverse chronological association can be found going from *Snack 23:21* to *Spare time/TV 22:19*. While it can be understood by looking at the data that it would be relatively

easy to predict that there is often a *Spare time/TV* activity before a *Snack* activity, it is much less clear that the former was done in preparation for the latter. This is because *Snack* occurrences are less easy to predict and thus it is less likely that they would be premeditated.

Indeed, if an action is planned beforehand, it must be planned from some information about previous events. If those previous events can be associated with the premeditated action, then the latter becomes easier to predict. Consequently, if an event type is not easy to predict, it may just be because it is not premeditated. However, it may also be because the data do not contain enough information about the events on which the premeditation is based. In short, how much information this reasoning really gives on the relation underlying a reverse chronological association is not clear.

If even the direction of the potential causality could not be determined, how could we claim to identify causal relations? This is one reason why we prefer to simply name *associations* the relations found during data interpretation. The task of inferring causal relations is left to the human reader of the generated text. For that task, humans have the advantage of being able to take into account a variety of knowledge and information sources.

After instances of association rules are identified, some extra associations are derived and added to the network. The *Repetition* association is generated whenever the type of activity that appears on the right side of the association rule also appears on the left side. The *Repetition* association is needed to communicate to the reader that the author (the computer) is aware that it is describing an event of the same type again. This confirms to the reader that the author is not just repeating the same statement emptily.

Conjunction is added when two sequential associations start or end at the same activity. Their other ends are then linked by a *Conjunction* association. This association groups events that may have something in common causally, such as having the same cause or same effect.

The *Instead* association appears when an unexpected sequential association is found. It indicates what would have been the most probable alternate activity according to the sequential association rule model. The *Instead* association is necessary to justify and explain to the reader the unexpected sequential association. Derived associations are shown on the right of the first column of Figure 2.

5 Narrative Generation

We now detail the workings of our model for the three stages of the standard NLG pipeline: document planning (composed here of event selection and document structuring), microplanning, and surface realization [27].

5.1 Event Selection

Event selection (step 3 of Figure 1) selects the events that are the most unusual, that is the least probable according to the sequential rules model.

```
1: procedure SELECT(event e)                    ▷ returns true if e is to be selected
2:     if ∃(x --p--> e) ∈ associations then      ▷ x is an event or a time
3:         for all x --p--> e ∈ associations do
4:             if p ≤ threshold then
5:                 return true
6:             end if
7:         end for
8:         return false
9:     else                ▷ y is an event, P(e|t_e) is the probability conditioned on time of e
10:        return  (∃(e --p--> y) ∈ associations, p ≤ threshold)  ∨
                   (P(e|t_e) ≤ threshold)
11:    end if
12: end procedure
```

Figure 5: Event selection algorithm.

Event selection has one parameter: a maximum probability threshold whose purpose is to generate reports that can vary in length depending on how unusual the period was. If an event is implied by another event or by the hour of the day, that is, if it is the second argument of a sequential association, it is selected if its confidence is lower than the threshold. If this is not the case, the event is selected if it implies another event with a probability lower or equal to the threshold or if it has a probability conditioned on time lower or equal to the threshold. The selection algorithm is formalized in Figure 5.

The purpose of the **if** block on lines 2–8 is to take advantage of the sequential association rules to retain only the events that are harder to infer from the others. Those events are considered more unusual. It is assumed that the reader is familiar with what usually happens as captured by the sequential association rules. The events that are rejected here are implied with a certain confidence from other events.

The logical disjunction on line 10 takes care of two more cases. First, it makes sure that associations that justified the selection of an event in lines 2–8 also have their other argument selected. This way those associations can be included in the document structure. The second part of the disjunction concerns events that are not implied by any other event or time. Those events are selected based only on their probability conditioned on time. This means that, in that case, an event is selected if events of that type rarely happen in the same temporal category (e.g. hour of the day) than that event.

For the purpose of event selection, an *Instead* association is considered an extension of the corresponding unexpected sequential association and uses the same probability.

Generally the ideal value of the maximum probability threshold varies according to how well the sequential rule model captures what usually happens and the desired average length of the generated text. In our case, this value was determined empirically by looking at several sample generated texts. Different values were tried to get texts which on average selected out enough data to be considered short stories while still displaying interesting

textual phenomena. Figure 2 shows the selected events for three maximum probability threshold values: 0.2 (underlined), 0.3 (in bold), and 0.4 (in italics). In this example, out of a total of 31 activities, 3 activities are selected with a threshold of 0.2, 10 with 0.3, and 17 with 0.4. The value used for the maximum probability threshold for the remaining of the running example is 0.3.

If they have not already been selected, the first and last events of the period of the report will be added to the selected events to become the initial and final situations of the connecting associative thread.

5.2 Document Structuring

Document structuring (step 4 in Figure 1) outputs a detailed plan of the overall structure of the narrative where only local decisions will need to be taken by the following stage of microplanning.

5.2.1 Connecting Associative Thread

The main idea behind the connecting associative thread is to give the text a simple narrative structure including a beginning, an ending, and a middle section that smoothly connects them. The importance of this structure for temporal data-to-text process was highlighted by a comparison with human written texts [20]. The connecting associative thread, as its name suggests, must also connect all the previously selected events with appropriate associations, so that the events form, as much as possible, a coherent whole for the reader.

The first event of the period (chronologically) is selected to be the beginning of the text and is called the initial situation (*Sleeping 00:33* in the example of Figure 2). The last event of the period is correspondingly called the final situation (*Sleeping 01:50* in the example). The (rest of the) selected associative sub-network will form the middle section (in bold type in Figure 2). The best event pairs are then chosen to link the selected events with each other. In the example, event pairs with sequential associations are preferred over those with only *Same category* associations. Manually set parameters, called association preferences, define in what order association types are preferred. They take a value between 0.0 and 1.0. A smaller value gives an event pair with this association type more chances to be chosen. When no other association is present, the default association of temporal proximity is used with association preference 1.0.

The association preference is combined (by averaging) with the relative temporal distance in order to favour temporally close event pairs. The relative temporal distance is the time elapsed between the end of the first event and the beginning of the second one divided by the total duration of the period. Averaging is used to combine the two because it preserves the range of values (in contrast with a simple sum) and is linear (in contrast with multiplication, for example). The linearity makes it easier to understand the impact of increasing or decreasing a parameter.

The resulting score is then used as a distance to compute a minimum spanning tree on the selected associative sub-network. However, there is one additional constraint: the final

situation must be a leaf. This is so it can be ordered last in a chain of associations in the text.

This minimum spanning tree is converted into a directed rooted tree by designating the initial situation as its root. This tree is hereafter called the *connecting associative thread*. The path from the initial situation to the final situation is the main associative thread. The other branches of the spanning tree are said to be dead-end threads because once the text has reached their end, it must go back to the connection point with the main thread before continuing towards the final situation. The connecting associative thread connects every event through the main thread and the dead-end threads.

The connecting associative thread is traversed in a specific way to obtain the order in which events will be mentioned in the text. Starting from the root, depth first traversal is employed with the addition of one constraint. The traversal is done so that a node that is part of the main associative thread is always visited last when the current node has more than one child. This has the desired effect that the traversal always begins with the initial situation and ends with the final situation. Recall that the final situation is always a leaf because of the constraint put on the spanning tree.

In the present algorithm, it is assumed that the best candidate events for the roles of initial and final situations are the chronologically first and last events, respectively. Consequently, the correct determination of the initial and final situations relies completely on an adequate definition of the type of the period of the narration. This choice must depend on the kind of input data. In the case of the ADL example, getting up in the morning and going to bed at night are appropriate choices of initial and final situations, because it is natural for humans to segment time in daily sleep/wake cycles.

5.2.2 Paragraph and Sentence Segmentation and Sentence Plan Assembly

In this step, the structured document content is segmented into sentences and paragraphs. There is not only one correct way to separate a narrative text into paragraphs [30]. Accordingly, some stylistic variation in segmentation is enabled by the model by adjusting two parameters: the target average number of events introduced in one sentence and the target average number of sentences in one paragraph. Those parameters are used to calculate the number of breaks needed between sentences and paragraphs. The candidate break points are between consecutive event pairs in the order given by the traversal of the connecting associative thread, as specified in the previous section. The actual break points are selected according to the distance computed previously for the determination of the minimum spanning tree. The greatest distances correspond to paragraph breaks, then sentence breaks, and lastly phrase boundaries. The exact distance values do not matter, as this operation does not rely on fixed distance thresholds. Paragraphs are boxed in Figure 6.

Between sentence breaks, consecutive event-describing clauses are grouped recursively by two to form longer phrases. Microplanning will later determine if the clauses are to be coordinated or one subordinated to the other. This grouping is done in order of increasing distance between the last event mentioned in the first phrase and the first event mentioned in the second phrase. The resulting binary tree constitutes the plan of that sentence. It will

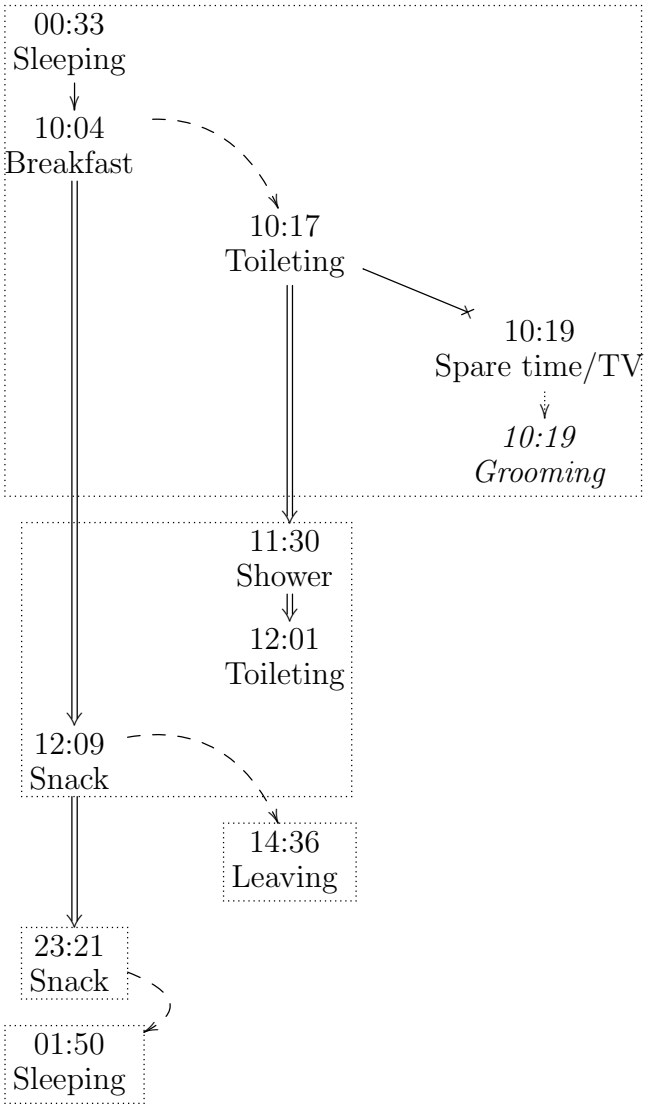


Figure 6: Connecting associative thread for user B on November 24, 2012. Arrows represent associations: simple: expected sequence; X-headed: unexpected sequence; double: *Same category*; dotted: *Instead*; curved and dashed: temporal proximity. Paragraphs are boxed. The vertical order of presentation is the order of mention in the generated text (Figure 7). For event selection, the maximum probability threshold was set to 0.3.

be used in microplanning to determine syntactic relations between phrases.

5.2.3 Mapping from Associations to Rhetorical Relations

At this point, a mapping is made between the selected associations and the rhetorical relations that will be expressed in the text. This mapping plays a role in the generation process similar to the logico-semantic relation to rhetorical relation mapping of Bouayad-Agha *et al.* [4]. The many-to-many mapping means that, in the proposed narrative generation model, how associations are detected is in principle independent of how they are expressed. (No claim is made here about how humans really identify relations and express them textually.)

Although the document structure is not a rhetorical structure in the sense of RST [19], the concept of rhetorical relation we use is similar. Here the purpose of the rhetorical relations is to indicate what family of linguistic means should be used to link the event descriptions in the subsequent microplanning stage to form a coherent discourse. Like in RST, they can either have an asymmetric nucleus-satellite structure or be multinuclear (more than one nucleus and no satellite). In asymmetric relations, the satellite is less central to the meaning expressed and could be dropped without rendering the nucleus incomprehensible. The reverse would not be true. In the case of multinuclear rhetorical relations, nuclei are interchangeable.

Table 2 shows how the mapping from associations to rhetorical relations is made in the example. To most associations corresponds one rhetorical relation. For the unexpected sequential association, two rhetorical relations need to be expressed: SEQUENCE and CONTRAST. The *Conjunction* and *Same category* associations both are expressed by a CONJUNCTION rhetorical relation, because *Same category* groups events that could play the same role at some level. Microplanning will not try to express any specific rhetorical relation in the case of the temporal proximity default association.

Only the *Instead* and *Repetition* associations correspond to nucleus-satellite rhetorical relations. Their second argument becomes the satellite of the corresponding rhetorical relation.

5.3 Microplanning, realization, and interpretation

Microplanning (step 5 of Figure 1) translates the rhetorical structure into a lexico-syntactic specification. Each sentence plan tree is traversed depth first. When a leaf is visited, a specification of the corresponding eventuality's description is produced from lexico-syntactic templates. When an internal node is visited, the rhetorical relations linking the two children nodes are expressed with appropriate discourse markers. Those markers are then used to assemble the lexico-syntactic specifications obtained from the children nodes.

The marking of rhetorical relations between sentences is more complex. To make sure that the reader can make the correct link with the previously mentioned argument of the discourse marker, anaphora is sometimes employed. The basic theory used for determining when this is needed is presented here.

Association type	Rhetorical relation(s)	Satellite
Expected sequence	SEQUENCE	—
Unexpected sequence	SEQUENCE & CONTRAST	—
Instead	INSTEAD	2nd argument
Conjunction	CONJUNCTION	—
Repetition	REPETITION	2nd argument
Same category	CONJUNCTION	—
Temporal proximity (default)	—	—

Table 2: Rhetorical relation(s) and satellite for each association type. No satellite means that the relation is multinuclear. No rhetorical relation means that none is to be explicitly expressed in the text.

Theories of intersentential coreference such as Centering Theory [8] focus in practice mostly on entities denoted by noun phrases. The referring expressions that need to be targeted as antecedents in the ADL running example are events denoted by clauses. The concept of main event is here introduced to deal in a pragmatic manner with the notion of prominence in the comprehension of a text about events. It is not meant to solve this problem completely. Each sentence and paragraph has one of the events it expresses called its main event. Each sentence is also assigned a previously mentioned main event called preceding main event. For any given sentence, the preceding main event represents the most prominent event in the mind of the reader when looking for the implicit argument of an intersentential discourse marker. The main event of a sentence is the one expressed by its first (as encountered by the reader) independent clause. An independent clause is assumed to be more prominent outside the sentence than a subordinate clause. The main event of a paragraph is the main event of its first sentence, because the first sentence of a paragraph is an expected position for the presentation of the topic of that paragraph [30]. Although this is not incorporated in the current algorithm, alternative ideal positions for the topic of a paragraph would be its second, third or last sentence. Inside a paragraph, a sentence's preceding main event is the main event of the preceding sentence. Because a paragraph break signals a change of topic, the preceding main event of the first sentence of a paragraph is the main event of the preceding paragraph. Only the most prominent event of the last paragraph is assumed easily accessible in that case.

In the process of combining sentences, a discourse marker is placed at the front of a sentence to indicate its parent relation in the connecting associative thread. If its parent is the preceding main event, the marker appears alone. In that case, it is assumed that the antecedent is prominent enough to be retrieved without further clarification. If the parent event is not the preceding main event, an anaphoric expression is added that restates the parent event. This is meant to facilitate the retrieval of this argument of the discourse marker. For example, the parent of *Toileting 12:01* in Figure 6 is *Shower 11:30*. Since

```
OrdonezB   Saturday, November 24, 2012 12:33 AM  -
                     Sunday, November 25, 2012 09:24 AM
-----------------------------------------------------------------------
OrdonezB got up at 10:02 AM and then he ate his breakfast.
As usual at 10:17 AM he went to the toilet but then he unexpectedly spent
1 hour in the living room instead of grooming.

In addition to having gone to the toilet at 10:17 AM, he took a shower at
11:30 AM. Also at 12:01 PM he went to the toilet. Beside his 10:04 AM
breakfast, he had a snack at 12:09 PM.

At 2:36 PM he left for 1 hour.

In addition to his 12:09 PM snack, he had a snack at 11:21 PM.

As usual at 1:50 AM he went to bed.
```

Figure 7: English generated text example for user B on November 24, 2012. The maximum probability threshold was set to 0.3. The text only describe *interesting* events, this is why nothing is listed between 2:36 PM and 11:21 PM

it is the main event of the preceding sentence, no anaphor is added and we have just the marker `also` in the generated text (Figure 7). On the contrary, the parent of *Snack 12:09* is *Breakfast 10:04*. It is located in another paragraph. Consequently, the marker becomes `Beside his 10:04 PM breakfast`.

Surface realization (step 6 of Figure 1) was performed using the SimpleNLG-EnFr Java library [35]. During surface realization, the syntactic and lexical specifications are combined with the output language grammar and lexicon to generate formatted natural language text. The lexico-syntactic templates used in microplanning were written for both English and French output languages. In combination with SimpleNLG-EnFr, this enabled bilingual generation. We also randomized between alternative discourse markers.

An example of English generated text corresponding to the preceding figures is given in Figure 7, the French version can be found in Vaudry [34].

Finally, in step 7 of Figure 1 a human reader combines his world and domain knowledge with the generated text to construct a causal mental representation of the events. For that the reader can follow the connecting associative thread through the text while trying to infer possible causal relations.

We hypothesize that statistically identifying sequential associations is a useful pre-processing of the data for the purpose of determining causal relations. Association rules based on similarity could also be helpful because events that share some similarity sometimes have the same cause or effect. Other association rules based on specific causal patterns could also give useful hints.

For example, the fact that the clauses expressing *Sleeping 00:33* and *Breakfast 10:04* are

coordinated in the same sentence and linked by the temporal marker `then` could lead the reader to different conclusions depending on his knowledge. On the one hand, they could think that maybe the user was particularly hungry when he woke up that morning; they could ponder why. On the other hand, they could also ignore this sequence as just a random event.

Another example: the fact that *Snack 23:21* references *Snack 12:09* could make the reader conclude that maybe the user was often hungry on that day and maybe there was a common cause for that. Or the reader may ignore this, reasoning that *Snack 12:09* was probably in reality a *Lunch* activity. The point is that some of the associations can help the reader in forming causal hypotheses. The reader can later verify those, for example by asking the user. Moreover, those causal hypotheses can help the reader remember the content of the text.

6 Evaluation

This section presents the method and results of an intrinsic evaluation that was conducted in the goal of measuring the textual quality of the ADL reports. It also compares the presented evaluation with previous temporal data-to-text evaluations. The generated texts, evaluation forms, and judge answers of this evaluation are publicly available.[1]

6.1 Method

For the text quality evaluation, as for the running example, texts were generated using data from the UCI ADL Binary Dataset [22] as input for training and generation. Reports were generated from both user A and B data. Because of the small size of the dataset, all the data for a given user was used as training data for that user's reports. To assemble the evaluation corpus, a report was generated for the thirty-two complete days of the dataset. A day was defined heuristically as starting and ending with a *Sleeping* activity lasting more than one hour and starting at least sixteen hours after the start of the last separating activity.

The maximum probability threshold parameter of event selection was adjusted in order that texts for both users have comparable average length. The maximum probability threshold was thus set to 0.4 for user A and 0.3 for user B. User A's routine seems to be easier to capture by the sequential rule model than user B's. Hence the probability for user A's activities according to the model is generally higher than for user B's. Only the English version of the reports were evaluated. The generated reports include the one presented in Figure 7.

In a scenario where the reports to be generated already exist in hand-authored form and suit well their purpose, it makes a lot of sense to use them as a gold standard to evaluate automatically generated reports. However, when it is not the case, as for the ADL reports, in an intrinsic evaluation reports written by humans for the occasion can at best serve as a point of comparison, but not as a gold standard. Indeed, nothing indicates that they would

[1] `http://rali.iro.umontreal.ca/rali/en/text-quality-evaluation`

be better than the automatically generated ones before they have been tested by routine use. And an intrinsic evaluation will not allow that to be estimated. Consequently, no gold standard was available for this evaluation. As for a baseline, as no equivalent of the generated ADL reports exists, there was also none. Therefore only the generated texts are evaluated. Still, a comparison with human-written texts could have been interesting. However for that domain experts would have been needed.

Human judges were asked to rate generated narrative texts on the following criteria:

1. Overall: *What proportion of the text corresponds to how that kind of report should be in general?*
2. Style: *What proportion of the text is written in a style appropriate for that kind of report?*
3. Grammaticality: *What proportion of sentences are grammatically correct?*
4. Flow: *What proportion of sentences flow well from one to the next?*
5. Vocabulary: *What proportion of word choices are appropriate?*
6. Understandability: *What proportion of the text is perfectly understandable?*

The judges had to evaluate the whole texts on a 0 to 5 scale for those six criteria. 0 meant that this aspect of the text was bad all over, while 5 meant that it was perfect everywhere in the text. Participants could also leave comments at the end of their evaluation of each text.

A sample of a blank evaluation form can be found at the location given earlier.1 Each form repeats the same seven questions for each of four or five texts: one question for each criterion, in addition to a space reserved for comments. The evaluation forms were presented in English only.

Thirteen volunteers were recruited to be judges. None were experts of an ADL-related healthcare domain. Because the texts to be evaluated were in English, only people with at least an approached native ability in English were accepted as judges. Judges evaluated four to five generated texts each, so that twenty-eight texts were evaluated by two judges each and four texts by one judge. The texts from the beginning, middle and end of each user sub-dataset were distributed evenly among evaluation forms. This way the reports from a given week were evaluated by different judges so that variation among judging styles were counterbalanced. The order was also alternated between forms in order to partially counterbalance possible order effects.

6.2 Results

If all the evaluations taken together are viewed as evaluating the data-to-text system as a whole, as opposed to individual texts, we get the results shown in Figure 8.

The best ratings are for Understandability and Vocabulary with peaks at 5 and 4, respectively. This indicates that the generated texts made sense for the judges. It also reveals that

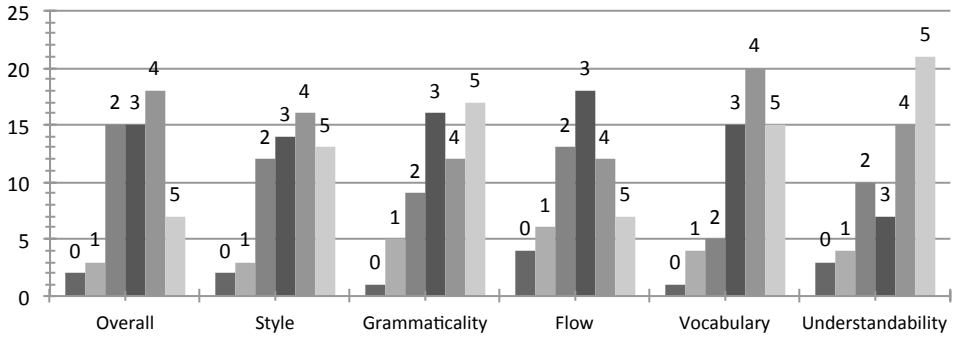


Figure 8: Results of the text quality evaluation. On the vertical axis is the number of evaluations and on the horizontal one are the ratings for each criterion.

the relatively simple mechanics behind lexical choice and lexical variation (lexico-syntactic templates and a little randomization between alternative discourse markers) were sufficient for the purpose of generating a report-style document.

The worst ratings are for Flow with a peak at 3. This could indicate some deficiencies in document planning and/or microplanning. However, according to the good Understandability ratings, the texts do not seem as badly planned as to be confusing. In document planning, the algorithm that determines the connecting associative thread could be revised. In microplanning, the computation of the preceding main event could be made to take into account more complex cases.

The results for Grammaticality are hard to interpret, since there are two peaks: one at 3 and one at 5. After looking at the evaluations, it seems to be because this criterion was not defined clearly enough. The same text could be rated very differently depending on the evaluator. Some judges seemed to classify as grammatical mistakes what others could consider merely stylistic peculiarities. For example, the two judges who evaluated text B20 gave ratings of 4 and 2 for Grammaticality. The former did not leave any comment on their evaluation. The latter commented: 'Commas should be used to set off introductory elements in most sentences.' This opinion could very well have influenced grammaticality judgments.

Overall and Style have most ratings ranging from 2 to 5, with peaks at 4. There seems to be a little more variation between evaluations in those criteria. Although they get better ratings than Flow, there is some room left for improvement.

6.3 Comparison with Previous Temporal Data-to-text Evaluations

In order to put the text quality evaluation just described in perspective, this subsection presents a comparison between it and other evaluations of previous temporal data-to-text systems.

Several types of method have been used to evaluate temporal data-to-text systems. Some only benefited from user comments at the time of designing the system [10]. Some used human ratings, similarly to the evaluation presented in this paper, to evaluate text quality. Wanner *et al.* [36] evaluated in this way sixteen air quality bulletins generated in each of

five languages. The results indicate that Comprehensibility, Fluency, Content relevance, and Level of detail needed improvement. Bouayad-Agha *et al.* [4] had fifty-one generated football match reports rated with good results for both Intelligibility/Grammaticality and Fluidity. Note that the number of rated generated texts in those studies and this paper are of the same order.

Since football match round-ups are regularly written by journalists and that the corresponding data are available, Bouayad-Agha *et al.* [4] also compared content selection between human and generated reports for thirty-six matches. They concluded that recall was good, but precision needed improvement. In this work, content selection evaluation could not be performed because human-written texts for the same data were not available. However, should this change, this would be an interesting evaluation to carry out.

Other evaluations necessitated experts to be performed. Those could be experts in the topic covered by the generated texts or linguistic experts. An example of the latter is the discourse analysis carried out in the Babytalk BT-45 project on three generated texts and three human written ones [20]. This analysis uncovered problems in narrative structure and in expressing temporal information. For the same project, an off-ward decision experiment was also conducted [26]. In this experiment, medical experts had to determine the required action based on historical data on twenty-four scenarios presented in one of three formats: graphical, generated text, or human expert-written text. The first two were found to perform equally well and were both inferior to the human-written text. Hunter *et al.* [13] describe an on-ward evaluation as part of the BT-Nurse project in which 148 generated texts were rated by nurses. Ninety per cent of ratings stated that the text was understandable, seventy per cent accurate, and fifty-nine per cent useful. Lastly, three experts and two members of the target audience rated three versions of a text for the Tag2Blog project [25]. The first generated version contained only indications about the spatial movements of a tagged bird. The second generated version had ecological insights automatically added. The third version was a human expert-post-edited version of the first generated one, also adding ecological insights. The best ratings were for the third, then the second versions. Compared with the evaluation presented in this paper, evaluations with experts require many more resources. Considering that such resources were not available as part of this research, an evaluation that did not require experts can be considered adequate.

As for comparing results with previous evaluations, unfortunately this is not possible. Even with evaluations that have similar methodologies (naive rating of texts), such as Wanner *et al.* [36] and Bouayad-Agha *et al.* [4], the differences are too great to enable direct comparison of results. Even if the evaluation criteria were the same, the fact that the texts are from different domains and generated from different datasets would be an obstacle.

The conclusion of this comparison with previous temporal data-to-text evaluations is that considering the limited resources, the methodology used is adequate. Although it is impossible to compare results directly with other evaluations, the presented text quality evaluation is sufficient to show that our approach is interesting and worthy of future research.

7 Conclusion

Contrary to other narrative data-to-text systems, we presented a model where document planning is centered around a mix of automatically mined and manually defined clues on causality. Those clues are called associations. The reader must use their world and domain knowledge to determine which of those associations are good hints toward identifying causal relations and which ones are red herrings. The generated text's communicative goal is to help the reader assimilate the facts necessary to construct a causal representation of the events. According to the model, the connecting associative thread allows the reader to follow associations from the beginning to the end of the text. This structure takes the form of a spanning tree over a selected associative sub-network.

The textual quality of the generated texts was rated by judges. The results show that the texts were understandable and the vocabulary adequate. However, flow between sentences, although not bad, could still be improved. A possible solution would be to modify document structuring such as to minimize discontinuities. According to the event-indexing model [37], sentence-reading times increase with the number of discontinuities in temporality, spatiality, protagonist, causality, or intentionality.

We have begun pilot testing a memorization experiment to test if the generated texts help the reader assimilate unusual facts independently of the domain. Apart from that, a task-oriented evaluation with domain experts could be organized.

Furthermore, texts could be generated from bigger ADL datasets, such as the CASAS datasets [5], or datasets belonging to other domains. The mining of association rules from multivariate and spatiotemporal data could also be explored. It would be interesting to fine-tune all parameters in each case to see if ideal values vary from one domain to another.

Subsequent experiments could investigate how, when the organization of the target text type can be rigorously analyzed, top-down constraints could be introduced in document planning while keeping a bottom-up approach for unconstrained parts.

Finally, how to determine the initial and final situations in every case is a problem not yet solved. In some cases, it can be desirable that the computer identify inside an arbitrarily determined stretch of time one or more event subsets likely to produce one or more suitable narratives. The associative network could be exploited to that end.

References

[1] Belén A Baez Miranda, Sybille Caffiau, Catherine Garbay, and François Portet. Task based model for récit generation from sensor data: an early experiment. In *5th International Workshop on Computational Models of Narrative*, pages 1–10, 2014.

[2] Belén A. Baez Miranda, Sybille Caffiau, Catherine Garbay, and François Portet. Generating Récit from Sensor Data: Evaluation of a Task Model for Story Planning and Preliminary Experiments with GPS Data. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 86–89, Brighton, UK, 2015. Association for Computational Linguistics.

[3] Mieke Bal. *Narratology : introduction to the theory of narrative*. University of Toronto Press, Toronto, 3rd ed. edition, 2009.

[4] Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, and Leo Wanner. Perspective-oriented Generation of Football Match Summaries: Old Tasks, New Challenges. *ACM Transaction on Speech and Language Processing*, 9(2):3:1–3:31, 2012.

[5] Diane J. Cook, Aaron S. Crandall, Brian L. Thomas, and Narayanan C. Krishnan. CASAS: A Smart Home in a Box. *Computer*, 46(7), 2013.

[6] Pablo Gervás. Composing narrative discourse for stories of many characters: A case study over a chess game. *Literary and Linguistic Computing*, 2014.

[7] Andrew S. Gordon. Commonsense Interpretation of Triangle Behavior. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[8] Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225, 1995.

[9] Przemysław Grzegorzewski. On the properties of probabilistic implications. In *Eurofuse 2011*, pages 67–78. Springer, 2011.

[10] Catalina Hallett. Multi-modal presentation of medical histories. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 80–89, 2008.

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

[12] Wilhelmiina Hämäläinen and Matti Nykänen. Efficient Discovery of Statistically Significant Association Rules. In *ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 203–212, 2008.

[13] James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial intelligence in medicine*, 2012.

[14] Anna Jordanous. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3):246–279, 2012.

[15] Harold H. Kelley. The processes of causal attribution. *American psychologist*, 28(2):107, 1973.

[16] Junseok Kwon and Kyoung Mu Lee. A unified framework for event summarization and rare event detection. In *CVPR*, pages 1266–1273, 2012.

[17] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

[18] L. Lee, R. Romano, and G. Stein. Introduction to the special section on video surveillance. *IEEE Transactions on pattern analysis and machine intelligence*, 8:740–745, 2000.

[19] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[20] A. McKinlay, C. McVittie, E. Reiter, Y. Freer, C. Sykes, and R. Logie. Design Issues for Socially Intelligent User Interfaces: A Discourse Analysis of a Data-to-text System for Summarizing Clinical Data. *Methods of Information in Medicine*, 49(4):379–387, 2009.

[21] James Niehaus and R. Michael Young. Cognitive models of discourse comprehension for narrative generation. *Literary and Linguistic Computing*, 29(4):561–582, 2014.

[22] Fco Javier Ordóñez, Paula de Toledo, and Araceli Sanchis. Activity Recognition Using Hybrid Generative/Discriminative Models on Home Environments Using Binary Sensors. *Sensors*, 13(5):5460–5477, 2013.

[23] Dana Mihaela Pavel. *MyRoR: Towards a Story-inspired Experience Platform for Lifestyle Management Scenarios*. PhD thesis, University of Essex, 2013.

[24] Alex Pentland, David Lazer, Devon Brewer, and Tracy Heibeck. Using reality mining to improve public health and medicine. *Stud Health Technol Inform*, 149:93–102, 2009.

[25] Kapila Ponnamperuma, Advaith Siddharthan, Cheng Zeng, Chris Mellish, and René van der Wal. Tag2blog: Narrative Generation from Satellite Tag Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 169–174, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[26] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816, 2009.

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

[28] Anne H. Schneider, Alasdair Mort, Chris Mellish, Ehud Reiter, Phil Wilson, and Pierre-Luc Vaudry. MIME-NLG in Pre-Hospital Care. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 152–156, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[29] Ivo Swartjes and Mariët Theune. A fabula model for emergent narrative. In *Technologies for Interactive Digital Storytelling and Entertainment*, pages 49–60. Springer, 2006.

[30] Douwe Terluin. From Fabula to Fabulous: Using Discourse Structure Relations to Separate Paragraphs in Automatically Generated Stories. Master's thesis, University of Groningen, 2008.

[31] Mariët Theune, Nanda Slabbers, and Feikje Hielkema. The Narrator: NLG for digital storytelling. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 109–112. Association for Computational Linguistics, 2007.

[32] Tom Trabasso and Paul van Den Broek. Causal Thinking and the Representation of Narrative Events. *Journal of Memory and Language*, 24(5):612–630, 1985.

[33] Tom Trabasso, Paul Van den Broek, and So Young Suh. Logical necessity and transitivity of causal relations in stories. *Discourse Processes*, 12(1):1–25, 1989.

[34] Pierre-Luc Vaudry. *Narrative Generation by Associative Network Extraction from Real-life Temporal Data*. PhD thesis, Université de Montréal, 2017.

[35] Pierre-Luc Vaudry and Guy Lapalme. Adapting SimpleNLG for bilingual English-French realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[36] Leo Wanner, Bernd Bohnet, Nadjet Bouayad-Agha, François Lareau, and Daniel Nicklaß. Marquis: Generation of User-Tailored Multilingual Air Quality Bulletins. *Applied Artificial Intelligence*, 24(10):914–952, 2010.

[37] Rolf A. Zwaan, Mark C. Langston, and Arthur C. Graesser. The Construction of Situation Models in Narrative Comprehension: An Event-Indexing Model. *Psychological Science*, 6(5):292–297, 1995.