Université de Montréal

Narrative Generation by Associative Network Extraction from Real-life Temporal Data

par
Pierre-Luc Vaudry

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Thèse présentée
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en informatique

Octobre 2016

RÉSUMÉ

Les données portant sur des événements abondent dans notre société technologique. Une façon intéressante de présenter des données temporelles réelles pour faciliter leur interprétation est un récit généré automatiquement. La compréhension de récits implique la construction d'un réseau causal par le lecteur. Les systèmes de data-to-text narratifs semblent reconnaître l'importance des relations causales. Cependant, celles-ci jouent un rôle secondaire dans leurs planificateurs de document et leur identification repose principalement sur des connaissances du domaine.

Cette thèse propose un modèle d'interprétation assistée de données temporelles par génération de récits structurés à l'aide d'un mélange de règles d'association automatiquement extraites et définies manuellement. Les associations suggèrent des hypothèses au lecteur qui peut ainsi construire plus facilement une représentation causale des événements. Ce modèle devrait être applicable à toutes les données temporelles répétitives, comprenant de préférence des actions ou activités, telles que les données d'activités de la vie quotidienne.

Les règles d'association séquentielles sont choisies en fonction des critères de confiance et de signification statistique tels que mesurés dans les données d'entraînement. Les règles d'association basées sur les connaissances du monde et du domaine exploitent la similitude d'un certain aspect d'une paire d'événements ou des patrons causaux difficiles à détecter statistiquement.

Pour interpréter une période à résumer déterminée, les paires d'événements pour lesquels une règle d'association s'applique sont associées et certaines associations supplémentaires sont dérivées pour former un réseau associatif.

L'étape la plus importante du pipeline de génération automatique de texte (GAT) est la planification du document, comprenant la sélection des événements et la structuration du document. Pour la sélection des événements, le modèle repose sur la confiance des associations séquentielles pour sélectionner les faits les plus inhabituels. L'hypothèse est qu'un événement qui est impliqué par un autre avec une probabilité relativement élevée peut être laissé implicite dans le texte.

La structure du récit est appelée le fil associatif ramifié, car il permet au lecteur de suivre

les associations du début à la fin du texte. Il prend la forme d'un arbre couvrant sur le sous-réseau associatif précédemment sélectionné. Les associations qu'il contient sont sélectionnées en fonction de préférences de type d'association et de la distance temporelle relative. Le fil associatif ramifié est ensuite segmenté en paragraphes, phrases et syntagmes et les associations sont converties en relations rhétoriques.

L'étape de microplanification définit des patrons lexico-syntaxiques décrivant chaque type d'événement. Lorsque deux descriptions d'événement doivent être assemblées dans la même phrase, un marqueur discursif exprimant la relation rhétorique spécifiée est employé. Un événement principal et un événement principal précédent sont déterminés pour chaque phrase. Lorsque le parent de l'événement principal dans le fil associatif n'est pas l'événement principal précédent, un anaphorique est ajouté au marqueur discursif frontal de la phrase.

La réalisation de surface peut être effectuée en anglais ou en français grâce à des spécifications lexico-syntaxiques bilingues et à la bibliothèque Java SimpleNLG-EnFr.

Les résultats d'une évaluation de la qualité textuelle montrent que les textes sont compréhensibles et les choix lexicaux adéquats.

Mots clés: Génération automatique de textes, data-to-text, planification de document, récit, données temporelles, relation causale, règle d'association, séquence, événement, connaissances du domaine.

ABSTRACT

Data about events abounds in our technological society. An attractive way of presenting real-life temporal data to facilitate its interpretation is an automatically generated narrative. Narrative comprehension involves the construction of a causal network by the reader. Narrative data-to-text systems seem to acknowledge causal relations as important. However, they play a secondary role in their document planners and their identification relies mostly on domain knowledge.

This thesis 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 associations suggest causal hypotheses to the reader who can thus construct more easily a causal representation of the events. This model should be applicable to any repetitive temporal data, preferably including actions or activities, such as Activity of Daily Living (ADL) data.

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 interpret a specific period to summarize, pairs of events for which an association rule applies are associated. Some extra associations are then derived. Together the events and associations form an associative network.

The most important step of the Natural Language Generation (NLG) pipeline is document planning, comprising event selection and document structuring. For event selection, the model relies on the confidence of sequential associations to select the most unusual facts. The assumption is that an event that is implied by another one with a relatively high probability may be left implicit in the text.

The structure of the narrative is called the connecting associative thread because it allows the reader to follow associations from the beginning to the end of the text. It takes the form of a spanning tree over the previously selected associative sub-network. The associations it contains are selected based on association type preferences and relative temporal distance. The

connecting associative thread is then segmented into paragraphs, sentences, and phrases and the associations are translated to rhetorical relations.

The microplanning step defines lexico-syntactic templates describing each event type. When two event descriptions need to be assembled in the same sentence, a discourse marker expressing the specified rhetorical relation is employed. A main event and a preceding main event are determined for each sentence. When the associative thread parent of the main event is not the preceding main event, an anaphor is added to the sentence front discourse marker.

Surface realization can be performed in English or French thanks to bilingual lexico-syntactic specifications and the SimpleNLG-EnFr Java library.

The results of a textual quality evaluation show that the texts are understandable and the lexical choices adequate.

Keywords: Natural language generation, data-to-text, document planning, narrative, temporal data, causal relation, association rule, sequence, event, domain knowledge.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF INITIALISMS AND ACRONYMS

ADL	**a**ctivity of **d**aily **l**iving
NICU	**n**eonatal **i**ntensive **c**are **u**nit
NLG	**n**atural **l**anguage **g**eneration
RST	**R**hetorical **S**tructure **T**heory
Triangle-COPA	**Triangle C**hoice **o**f **P**lausible **A**lterntives

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my supervisor Guy Lapalme for his precious support, beginning with the opportunity he gave me as an undergraduate to intern in his lab, and through all my doctoral studies and research. His wisdom and pragmatism helped me avoid straying from the path and reach my goal.

Very friendly thanks to analyst Fabrizio Gotti and to everyone at the laboratoire de recherche appliquée en linguistique informatique (RALI) for the stimulating discussions and moral support.

My grateful thanks are extended to the Fonds de recherche du Québec - Nature et technologies (FRQNT) and the Département d'informatique et recherche opérationnelle (DIRO) de l'Université de Montréal for their financial support.

Last but not the least, I would like to thank my partner Vanessa, my parents Doris and François, and my sister Myriam for their patience and loving support during this challenging period of my life.

CHAPTER 1

INTRODUCTION

When I was a child, like many others, when I came back from school my parents used to ask me something like: "What did you do at school today?" Then I would tell them about anything of note that had happened on that day. What went on in my mind between the question and the answer, I would be very curious to know. How were all the innumerable perceptions and thoughts of one day processed, summarized, and converted into language? However, this is not just a matter of curiosity. Our technological world is in great need of a solution to a very similar problem.

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 clickthrough statistics. These examples have in common large volumes of frequently updated non-textual event data. Much of this temporal data is 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. Indeed, the data must not only be described, but also analyzed, summarized, and explained. One way of presenting voluminous temporal data is to use Natural Language Generation (NLG) technology to produce a narrative text summarizing the events of a given period.

Experiments have shown that a narrative written by a domain expert can be a better support for decision-making than a graphical presentation of the same data (Law et al., 2005). Unfortunately automatically generated narratives do not yet achieve the same level of performance (Portet et al., 2009). Experts in discourse analysis have concluded that the problem may lay in the narrative structure: deficiencies in narrative flow and narrative details impacted negatively on coherence (McKinlay et al., 2009).

Hints on how to solve this problem may be found in cognitive psychology. According to experiments, narrative comprehension in humans seems to involve the construction of a causal network (Trabasso and van Den Broek, 1985, Trabasso et al., 1989). It is only natural, then, that causal networks have been applied to the automatic creation of fairy tales (Swartjes and Theune, 2006, Theune et al., 2007). Moreover, several narrative data-to-text systems already identify and make use of some causal relations (Bouayad-Agha et al., 2012, Hallett et al., 2006, Hunter et al., 2012, Wanner et al., 2010). Going further, in Vaudry and Lapalme (2015) we have tried to extract a form of causal network from temporal data and use it to build the structure of the generated narrative. We used data mining techniques to extract sequential association rules and interpreted them as indicating potential, approximate causal relations. The resulting causal network was used to express locally some rhetorical relations in the sense of the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). However, we did not succeed at that point at exploiting it to build a complete document structure that would give the text a global coherence.

The next year, we proposed an assisted temporal data interpretation model (Vaudry and Lapalme, 2016) that I describe in more detail in this thesis. This model uses a combination of automatically mined sequential association rules and a limited number of manually written world and domain knowledge rules to extract an associative network from temporal data. From this is built a document structure linking together the most unusual events. An associative thread can be followed by the reader from the beginning to the end of the generated narrative. Finally, this intelligent preprocessing assists the human reader in applying their world and domain knowledge to form a causal representation of the events.

The thesis is organized as follows. After presenting the previous work that serves as theoretical foundations for this thesis (Chapter 2), an entire chapter is dedicated to presenting the assisted temporal data interpretation model (Chapter 3). Next its components are detailed, beginning in Chapter 4 with those that deal with extracting an associative network from the data. Chapter 5 completes the model 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 Chapter 6.

CHAPTER 2

PREVIOUS WORK

This chapter presents various work that, taken together, constitute the theoretical foundations of this thesis. It begins with what seems obvious by resorting to the discipline of narratology for a definition of the concept of narrative. It proceeds with psychological notions about causality and an overview of the role of narratives in scientific explanations, before a short excursion in the world of data mining. Last but not least, several research projects in the area of natural language generation are summarized, be they about creating fictional narratives or generating from real-life data.

2.1 Narratology

To make sure that a narrative text is an appropriate format for data-to-text from temporal data and to begin to understand how to generate one, let us first turn to the field dedicated to its study, that is narratology.

Genette (2007, p. 13) describes three senses of the word *narrative* (*récit* in French). The first sense is a text or speech telling about one event or a sequence of events. The second is a sequence of events and their relations (sequence, opposition, repetition, etc.) being the object of the narrative in the first sense. The third is an act of telling a story. The first definition confirms that a narrative expresses temporal data. The second one adds that the relations holding between the events are of importance.

Bal (2009, p. 5) adds to the first sense, that she calls *narrative text*, more media such as imagery and sounds. Her *story* would correspond to the second sense. The novel element here is that the events must be caused or experienced by *actors*. Those are agents that are not necessarily human but have the capacity to perceive and act. This definition would mean that temporal data involving actors would make better material for narrative generation than temporal data concerning only inanimate objects.

2.2 Psychology

Psychology is very useful as part of this thesis for the insights it gives on how humans identify causal relations in real-life and in the process of comprehending narratives.

2.2.1 Causal Attribution in Everyday Life

Attribution theory (Kelley, 1973) deals with how people make causal explanations in everyday life. It models what kind of information people use to make causal inferences and how they use it. It has first developed to study how people choose between an internal and an external cause: is this behavior explained by the personality of the person doing it (internal) or the situation in which it happens (external)?

One useful concept of attribution theory for analyzing temporal data is the *covariation principle*. This states that when selecting among possible causes, people tend to choose the one that covaries over time with the potential effect. It is also stated that this principle implies temporal contiguity between the cause and the effect. For the subject to be able to observe covariation, there must be occurrences of the potentially causally related pair appearing in close temporal proximity and other instances when neither occurs.

2.2.2 Narrative Comprehension

Trabasso and van Den Broek (1985) and Trabasso et al. (1989) used the concepts of *causal network* and *causal chain* to explain the process of narrative comprehension in humans. Those causal networks are essentially composed of physical and mental events and states (of which *actions* and *goals*) connected by causal relations. Restrictions apply on which types of causal relation can connect which types of event or state. Goals *motivate* actions and mental reactions are *psychologically* caused. Actions and other physical events can *physically* cause other physical events. If an event or state is one of several causes but not a sufficient cause, it is a case of *enablement*. The causal chain comprises the events that are on a path that traverses 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. This indicates that causal relations are at the heart of how humans process information submitted in the form of a narrative.

Zwaan et al. (1995) present the event-indexing model, a model of how readers mentally reconstruct the situations described by a narrative. According to this model, as each event mentioned in the text is comprehended, the current situation model is updated on five indices: temporality, spatiality, protagonist, causality, and intentionality. It has been demonstrated that sentence-reading times increase with the number of indices that present a discontinuity. This shows that events are related and grouped in the text according to various dimensions of similarity.

Centering theory (Grosz et al., 1995) models the focus of attention while reading a text. From that it makes predictions on the relation between the choice of referring expression encountered and how coherent the text appears to the reader. Each utterance has a set of entities called *forward-looking centers*. One of them plays also the role of *backward-looking center*. The backward-looking center of an utterance points towards one of the forward-looking centers of the previous utterance. Forward-looking centers are partially ranked by prominence. Coherence and realization are affected by the rank and role of the backward-looking center in the previous utterance. Those general principles can inspire how to refer to events in the narrative.

Neuroscience experiments have shown that in the telling of a spoken narrative, several areas in the listener's brain are coupled with ones in the speaker's (Silbert et al., 2014). This brain-to-brain interaction highlights the importance of imitating natural narratives for effective temporal data-to-text.

2.3 History and Philosophy of Science

Section 2.2.1 briefly presented attribution theory. This theory assumes that the man on the street uses a naive version of the scientific approach used in behavioral science (Kelley, 1973, p. 109). Since data-to-text systems must often deal with scientific concepts in order to provide useful insights into data (such as in Hallett et al., 2006, Hunter et al., 2012, Ponnamperuma et al., 2013, Portet et al., 2009, Turner et al., 2008, Wanner et al., 2010), it is useful to consider

the role of narratives in scientific explanations.

Athearn (1994, pp. 58–59) explains that historically, causal stories have been an important aspect of the sciences. An exception is physics, which left this behind at the end of the nineteenth century to replace it with descriptions in the form of mathematical formulas. However, narratives remain essential to other sciences, such as geology and biology.

In political science, for example, Dowding (2015, pp. 94–95) states that scientific models can provide an explanation by modeling a mechanism, which he describes as "a narration that makes sense of the data." He adds that a characteristic of narratives is that they make a causal claim.

These thoughts about the history and philosophy of science tell us that a narrative is an appropriate way to scientifically model phenomena present in temporal data. They also highlight the importance of causality in those narratives.

2.4 Data Mining

Data mining techniques can be used to automatically extract knowledge about associations between events in a dataset. Hamalainen and Nykanen (2008) presents a method for finding statistically significant association rules. Traditional association rules are selected based on frequency and confidence. Given a rule candidate $X \rightarrow Y$ with X and Y being events, its frequency is $P(X,Y)$ and its confidence is $P(Y|X)$. Selecting based on statistical significance is desirable because associations between infrequent events are as relevant for the narrative as associations between frequent events. Significance can be measured by computing the probability of independence of the two parts of the rule. If that null hypothesis would be true, then $P(X,Y) = P(X)P(Y)$. The probability of independence is given by the binomial distribution because an event either occurs or does not occur. This is formalized in the following p-value formula. The lower this value, the more statistically significant a candidate rule is. $m(X)$ is the number of times X occurs, $m(X,Y)$ is the number of times X and Y occur together, and n the total number of occurrences in which X and Y could have occurred.

$$p(X \to Y) = \sum_{i=m(X,Y)}^{m(X)} \binom{n}{i} (P(X)P(Y))^i (1 - P(X)P(Y))^{n-i}$$

An association rule can be interpreted as a probabilistic implication. A probabilistic implication is one that holds with a certain probability (Grzegorzewski, 2011). In that case the probability of this implication is the confidence of the rule.

2.5 Natural Language Generation

As one might expect, most of the previous work presented in this chapter belongs to the area of natural language generation (NLG). Reiter and Dale (2000) describe an NLG architecture that is representative of a number of existing NLG systems. It consists of a pipeline composed of three components: the document planner, the microplanner, and the surface realizer. The task of the document planner is to determine the content of the document along with its structure. It usually requires application domain knowledge about what information is appropriate for the communicative goal and the user model. Knowledge about how documents are structured in that domain may also be required. The document planner outputs a document plan and passes it to the microplanner. The microplanner must make some decisions that depend more on language and effective writing and less on domain knowledge. For example, it may be left to the microplanner to determine exactly how to package the selected content into sentences and how to choose referring expressions. Purely linguistic knowledge about grammar and document structure is stored in the third component, the surface realizer. Its job is to convert an abstract document specification into a string. It may have to put words in the right order, find the correct verb inflection and any auxiliaries corresponding to a particular tense and produce the proper mark-up to signal paragraph boundaries. This pipeline architecture is the one adopted in this thesis.

According to Reiter and Dale (2000), many NLG systems structure document content using ideas from Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). RST is a theory about textual coherence. It posits a hierarchical document structure composed of recursive text segments. Two or more segments are grouped with a rhetorical relation to form a bigger segment. Some rhetorical relations, such as *Circumstance* and *Elaboration*, are binary and asymmetrical:

one of their arguments is the *nucleus* and the other the *satellite*. The nucleus is more important for the writer's purpose. The satellite could be substituted by another satellite of the same kind. Other rhetorical relations are multinuclear. All their arguments have the same status. *Contrast* and *Sequence* are examples of multinuclear relations.

An advantage of placing the tasks of surface realization in a separate component of the pipeline is that it can be reused in different NLG systems. The grammar and lexicon do not in principle have to be adapted much when changing domain but not language. SimpleNLG-EnFr (Vaudry and Lapalme, 2013) is a bilingual English-French surface realizer that is used as part of this thesis. It is an adaptation of the English surface realizer SimpleNLG (Gatt and Reiter, 2009). It can be used to construct lexico-syntactic specifications and linearize them while taking care of morphology and punctuation. It can also accept canned strings inserted anywhere in the structure.

Systems generating narratives fall into two broad categories: those that do so based on fictional data and those generating from real-life data. The two following subsections summarize a few of them. Another subsection covers NLG evaluation.

2.5.1 Fictional Narrative Generation

Swartjes and Theune (2006) and Theune et al. (2007) apply causal networks to the automatic creation of fairy tales. A multi-agent system simulates a story world where characters can have goals, emotions and perceptions and can perform actions. This simulation outputs a story representation in the form of a causal network inspired by Trabasso et al. (1989). This serves as input to the NLG component of their storytelling system. The document planner component prunes the causal network and transforms the causal relations into rhetorical relations to form a rhetorical structure. The rhetorical relations used include Cause, Contrast, Temporal and Additive. Microplanning then maps the leaves of the rhetorical structure to syntactic dependency trees. After lexicalizing the nodes of those trees excepting referring expressions, they are syntactically aggregated with cue phrases depending on the rhetorical relation. Then referring expressions are generated and the now fully lexicalized dependency trees are linearized by a surface realizer. However, the system as described in those papers does not yet handle contrast relation detection

and separation of the rhetorical structure into paragraphs. In the same research project, Terluin (2008, chap. 7-8) studies the paragraph segmentation problem. Experimental results show that there is not one unique correct way to segment a text into paragraphs. However, sets of sentences covering a single event or issue (possibly containing elaborations) are never separated.

With the help of focus and inferencing models, Niehaus and Young (2014) generate fictional narratives in which some events need to be causally inferred by the reader. For example, suppose the reader is told that Robbie waits in a dark alley with a gun for someone to come out of the bank. The reader is next told that Sally goes to the bank, but the next day is unable to buy the dress she wanted to buy. The reader should then infer that Robbie robbed Sally. This is because this inference is both enabled (the explanation is simple and there are few alternatives) and necessary (the reader must make the inference to make sense of the story). Those causal inferences must be precisely defined as part of the input in order to be modeled.

León and Gervás (2010) also use causality-related relations to structure narratives. Their algorithm learns preconditional rules between events of a fictional story with the help of human feedback. Existing stories are first translated by a human into an event sequence, with events represented by predicates. An assumption is made that a valid set of preconditional links is one in which there exists a directed path from any event to the last event of the story. In other words, every event must be directly or indirectly a precondition to the last event of the story. Although this may make sense for a fictional story, it could involve eliminating important information when starting from real-life data.

2.5.2 Narrative Data-to-text

This subsection gives an overview of several data-to-text systems generating narratives from temporal data. Most identify and make use of some causal relations. Additionally, note that to achieve this they tend to rely on domain-specific knowledge.

Hallett et al. (2006) describes a medical history visualization system that integrates graphics and text. A semantic network extracted from patient records serves as input to the generation process. It contains information on problems, investigations, and treatments and causal relations between them. Depending on the type of summary requested, a list of concepts essential to the

summary will constitute the *summary spine*. To each event of the summary spine, some related events may be added depending on the summary type and length. Events are then grouped by time period and/or event type. Within groups, events are structured with rhetorical relations that are either found in the input or deduced from domain specific rules. Those include Cause, Result, and Sequence. The List relation is used for unconnected events. Events part of the summary spine are placed in the role of a nucleus of a rhetorical relation and other events preferably as satellites.

The Babytalk project aims at producing textual summaries from continuous physiological signals and discrete events recorded in a Neonatal Intensive Care Unit (NICU). Portet et al. (2009) describes the BT-45 prototype which summarizes periods of about 45 minutes. It uses expert rules to determine if two events must be *linked* because they are causally or otherwise related. The document planner works similarly to that of Hallett et al. (2006). It first identifies as *key events* those that have the highest importance and creates a paragraph for each. In each paragraph it places first the key event itself, then linked events, and lastly events happening at the same time. Paragraphs are ordered first by body system and then by the start time of their key event.

Discourse analysts compare BT-45 generated texts with human-written texts in McKinlay et al. (2009). They note that the texts differ in terms of narrative structure. The beginning and end of the human-written narratives are clearly marked by a temporal marker and a description of the current state of affairs. This description only contains the information judged relevant with what happened in the middle section of the narrative. The information found at the end represents an update on what was first presented. The generated texts also present apparent inconsistencies and statements which do not seem topically relevant. The human-written texts are better at making apparent relations between events that presented those events as a coherent whole. They also tend to foreground the baby under care as an actor and provide information about his or her experience.

After learning from BT-45, the Babytalk project continues with BT-Nurse (Hunter et al., 2012). This system summarizes 12 hours of NICU data to support nurse shift handover. Document planning is realized using a mixture of techniques. The *Current Status* section is planned

with *schemas*, that is fixed structures filled with predetermined types of events. The *Events During the Shift* section is planned essentially as in BT-45. Some narrative deficiencies of BT-45 are addressed. The BT-Nurse document planner keeps track of the most recent information communicated about each data channel to avoid continuity problems. For example, if the information that *X increased to 10* and then *X increased to 8* is selected, to avoid an apparent inconsistency it is important to additionally mention that *X decreased to 2* between the two increases. This can take the form of a subordinated clause placed in front of the clause describing the second increase. Another important point is *temporal planning*. This is required because events inside a paragraph are not necessarily presented in chronological order. An event has an *event time* and a *reference time* that are used to determine tense. The former is the time at which the event happened. The latter can be equal to event time in a descriptive section, but can correspond to the time of the last mentioned event in a more narrative part. A perfect tense is used when the reference time is after the event time. Furthermore, to determine the event time, the temporal focus can be on the initial state, change of state, or final state corresponding to an event. Regarding temporal modifiers, the strategy is to add an absolute time phrase to all key events and to long events when the focus is on their end state. The authors conclude that more effort is still needed to generate narratives that make the big picture clear and express properly temporal, causal, and other discourse relations.

Turner et al. (2008) generate a weather forecast from spatio-temporal meteorological data. They identify spatial reference frames used by experts to partition space into meaningful sub areas for descriptive purposes: altitude, absolute direction, coastal proximity, and population. Data points are enriched with those characteristics with the help of external geographical information systems. The communicative purpose of spatial descriptions in that type of weather forecast is to explain each weather phenomenon by expressing a causal relation between it and some spatial reference frame. The evaluation highlights that forecasters would like to see more causal linking between events of a generated weather forecast in order to get a more fluent story.

The MARQUIS system generates user-tailored air quality bulletins for five European regions (Wanner et al., 2010). It takes as input numeric monitored and forecasted air pollutant and meteorological time series. It can indicate causal relations between meteorological conditions

and pollutant levels. For this purpose, depending on the region correlations are determined by multiple regression, machine learning, or expert rules models. The causal relations can be lexicalized in various ways depending on the syntactic context: noun (cause, consequence), verb (to cause), adjective (responsible, due), or adverb (because, due, therefore). This is made easier by the use of a surface realizer that starts from a conceptual representation instead of a syntactic one. This surface realizer is implemented with the graph transducer workbench MATE (Bohnet et al., 2000).

Bouayad-Agha et al. (2012) describe a system for generating perspective-oriented football match summaries. The small generated narrative is adapted to reflect the affinity of the targeted reader for one of the teams. The same surface realizer as MARQUIS is employed. *Logico-semantic relations* including causation are inferred between input concepts using domain-dependent rules. Logico-semantic relations can in principle be mapped to different rhetorical relations depending on the perspective chosen. For example, if a neutral perspective if desired, causation can be mapped to a temporal circumstance instead of a rhetorical cause or result. In practice perspective is communicated mainly through lexicalization. Semantic and lexical units are qualified with a polarity: positive, neutral, or negative. For example, the concept of causation can be expressed by three causation semantic units, one for each polarity. The semantic positive causation can in turn be expressed not only by the verb *cause* but also by the positive-sounding preposition *thanks to*, among others.

The Tag2Blog system generates narratives centering on satellite tagged wild animals (Ponnamperuma et al., 2013). The narrative places the animal's movements in its ecological context by associating its time-stamped locational information with data about habitat types, terrain features, place names, and weather conditions. This enriched data gives an impression of the animal's behavior: site fidelity and exploration, feeding and roosting, and social behavior. The needed expert knowledge is dependent on the species. The document plan is composed of two paragraphs each specified by a schema. The first one summarizes the overall movement pattern for the week. The second one, more detailed, describes interesting behavior during the week.

Gervás (2014) addresses the issue of planning a narrative about events involving several characters at different locations over the same period. The chosen experimental input data is

the description of a chess game in algebraic notation. The characters are the pieces and their perceptual range on the board is limited to a few squares. A *fiber* is defined as the sequence of events that are perceived by a given character. Once a set of fibers has been selected based on the communicative goal, *splicing* can begin. This operation consists of identifying potential breakpoints in each fiber, together with pairing breakpoints so that appropriate transitions between fiber fragments are obtained. A sequence of fiber fragments is called a *yarn*. For the initial version of the model, the splicer begins with the fiber that starts earliest, ends the fragment at a random point, and continues with the fiber that has been less fragmented until all the selected fibers are consumed. To indicate any change of context between fiber fragments now adjacent in the yarn, contextualization devices are placed at the beginning of each fragment: temporal expressions or discourse markers, spatial expressions, tense, setting or character attitude indications, etc. The yarn is then converted to a *narrative span* by removing redundant information in successive event descriptions. According to the author, further work should integrate causal relations along the lines of Trabasso et al. (1989). The detection of causal relations between events in different fibers, in particular, could guide better splicing into yarns.

Baez Miranda et al. (2014, 2015) generate a narrative from sensor data acquired during a mountain ski excursion. A domain-specific *task model* provides top-down constraints on the sequence of scenes that can be identified in the data to form the structure of the narrative. The task model organizes activities into a hierarchy of tasks and sub-tasks. Each task is characterized by an actor and objects in the actor's environment. The model can represent temporal constraints, dependencies and causal relations between tasks. Based on horizontal and vertical speeds, GPS data is divided into segments classified as *ascending*, *moving forward*, *descending*, or *break*. The segments are then incorporated into an ontology encoding the task model and enriched with point-of-interest data. The narrative's content is selected based on what is valid according to the task model. The segments containing the beginning, end, and goal of the excursion are always included.

Farrell et al. (2015) describe a model for the generation of narrative explanations from error trace data. The model uses finite-state technology. Trace explanations are specified for a specific domain by associating error trace regular expressions with natural language strings. Among

other things, this can be used to recognize and express the possible causes of an error.

Ahn et al. (2016) generate small narratives from causal graphs. Those causal graphs are abducted from first order logic descriptions of the actions of simple imaginary characters taken from a commonsense inference benchmark (Gordon, 2016). To this end, commonsense knowledge of actions, social relationships, intentions, and emotions are manually encoded. To generate more fluent narratives, paraphrases are first generated by applying rule-based grammatical transformations. Then a probabilistic parser is used to select the best one.

2.5.3 Evaluation

This subsection summarizes selected research that is relevant for narrative data-to-text evaluation. The ideas presented here inspired the evaluation methods described in Chapter 6.

Belz and Reiter (2006) compares automatic and manual methods for the evaluation of knowledge-based and statistical NLG systems. They classify evaluation methods as either intrinsic or extrinsic. The former consist of evaluating the generated texts themselves, either by asking subjects to rate them or by automatically comparing them to reference texts using evaluation metrics. The latter methods evaluate the generated texts by measuring their impact on task performance, counting expert post-edits, or measuring reading times.

Jordanous (2012) proposes a standardized procedure for evaluating computational creativity systems. As part of this work, a corpus of academic papers was analyzed to find the words most significantly associated with creativity. Those words were clustered and 14 components of creativity identified. Those that are the most relevant for the model proposed in this thesis are: dealing with uncertainty, domain competence, and originality. In particular, note that originality can mean to show previously unassociated concepts as linked with or without specifying how they are related.

Callaway and Lester (2002) designed a narrative prose generation system that takes as input a high-level story specification. It was subject to an intrinsic evaluation by presenting the output of different versions of the system to subjects who rated it on 9 criteria. Some pertained more to the formal properties of the generated texts: *style*, *grammaticality*, *flow* (between sentences), *diction* (word choices), and *readability*. Others were more about the semantics of the texts: *logicality*

(omissions, being out of order), *detail* (too much or too little), and *believability* (behavior of the characters). There was also an *overall* criterion which asked how good the generated stories were compared to fairy tales in general.

Portet et al. (2009) conducted a task-based extrinsic evaluation of the BT-45 system. Doctors and nurses were given a human-written text, a generated text or a graphical presentation of historical data about a baby that had been hospitalized in the past. They then had to make decisions about the action(s) to take. A score was computed which rewarded appropriate actions and penalized inappropriate ones. The human-written texts proved superior to the other two. There was no significant difference between generated texts and graphical presentations.

Hunter et al. (2012) evaluated their BT-Nurse system by generating texts for currently hospitalized babies at shift handovers and asking the incoming and outgoing nurses to rate them. The evaluation criteria were *understandability* of the text, *accuracy* of the content, and *helpfulness* in writing a shift summary (outgoing nurse) or in planning care (incoming nurse). A significant majority of ratings were positive for each criterion. This evaluation could be classified as intrinsic because the text themselves are rated by experts. However, helpfulness could be considered an extrinsic criterion because it relates to what the nurse can do with the text, i.e. its function.

2.6 Conclusion

This chapter presented some necessary basic notions: definitions of narrative, the covariation principle, causal networks and causal chain, narratives in science, association rule mining, an NLG pipeline architecture, Rhetorical Structure Theory, various examples of NLG systems, and intrinsic and extrinsic evaluations. It can be observed that the notion of causality is associated with narratives and temporal data in narratology, psychology, fictional narrative generation, and temporal data-to-text. 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 chapter presents a narrative data-to-text model that addresses those issues.

CHAPTER 3

ASSISTED TEMPORAL DATA INTERPRETATION MODEL

This chapter proposes a model showing how automatically extracted and manually written association rules can be combined to build the structure of a narrative from real-life temporal data. First the main characteristics of the model are presented. Then the communicative goal of the generated text in the context of this model is explained. Follows a brief overview of the main steps included in the model. Finally possible applications are discussed and an example application in the Activity of Daily Living domain is introduced.

3.1 Main Characteristics of the Model

The model presented in this thesis 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. All those characteristics are elaborated in the following subsections.

3.1.1 A Narrative Generation Model

A narrative is a text presenting with a certain angle a series of logically and chronologically related events caused or experienced by actors (Bal, 2009, p. 5). Therefore, this model assumes that generating a narrative is a natural way to communicate temporal data including actions or activities, if that corresponds to the needs of the users. To generate a text that readers recognize as a good narrative, this model takes inspiration from narrative comprehension theory (Trabasso and van Den Broek, 1985, Trabasso et al., 1989). More precisely, it assumes that the process of narrative comprehension involves the mental construction of a causal network by the reader.

3.1.2 A Data-driven Model

The model includes the extraction of sequential association rules using data-mining techniques (Hamalainen and Nykanen, 2008). This means that patterns where it does not seem plau-

sible that chance 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 particular dataset.

To maximize the scope of this thesis, it was important that the model be applicable to a wide range of domains. That is why a domain-independent bottom-up approach to document planning was chosen. That said, given knowledge about the structure of the narratives needed for a specific application, it could plausibly benefit from the addition of top-down specifications. For example, Baez Miranda et al. (2014, 2015) use a task model providing constraints on the sequence of scenes that can be identified in ski excursion data to structure the narrative. Top-down specifications of this kind are not addressed in this thesis.

3.1.3 An Assisted Interpretation Model

What makes this model an assisted interpretation model instead of a fully automated one is its recognition of the need of incorporating human input. Associations are provided that the reader must interpret to arrive at a causal explanation. In addition, world and domain knowledge are added manually into the system.

3.1.3.1 Associations, Not Causal Relations

In the course of my research, I found that it was very difficult to infer even the direction of a potential causal relation from an extracted sequential association. The initial hypothesis was that approximate causal relations could be derived from sequential association rules. For this purpose it was assumed that temporal precedence was an indicator of potential causality. Sequential association rules were thus mined with a chronological direction only.

A first paper was published (Vaudry and Lapalme, 2015) and the claim of identifying even approximate causal relations was criticized in the anonymous reviews. Some argued that domain knowledge was necessary. This I agreed with. I eventually integrated the possibility of having manually written association rules that could reflect world and domain knowledge.

A later hypothesis suggested mining reverse chronological sequential association rules to try to capture underlying goals having a later manifestation. If for an action sequence AB, it is possible to predict the previous occurrence of A given the later occurrence of B, the hypothesis was that A was possibly done in order to be able to do B. In other words, the goal of A could be the enabling of B. Reverse chronological association rule mining was implemented to test this. After looking through the results, I came to the conclusion that there was no clear link between the direction of the rule and the direction of a potential causal relation.

Because of those considerations, I no longer claim to identify causal relations, even "approximate" ones. The relations found are simply termed *associations*. By association, I 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 formal similarity. For the purpose of narrative comprehension, I assume that interesting associations are those that can help the reader formulate causal hypotheses.

In contrast with work on reader inference in fictional narrative generation such as Niehaus and Young (2014), 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 unassociated events. In this sense and to the extent that it accomplishes this, it can be considered to produce original, creative text (Jordanous, 2012, p. 257).

3.1.3.2 Added World and Domain Knowledge

World and domain knowledge can be formalized as various forms of association rules. Only a limited number of rules is necessary. Those rules can be manually entered or come from an existing ontology, for example. Belonging to the same ontology class can be made into a type of association. Events of the same class can have similar causes and effects. Other possibilities

include creating associations between events having the same location (or locations of the same class) or sharing an argument (or arguments of the same class). Generally, any knowledge about possible causality that can be formalized as a rule can be used here.

3.1.3.3 Construction of a Causal Representation by the Human Reader

In this model, the most important associations found between selected events form a chain of associations throughout the text. The model assumes that this associative thread can be followed by the human reader from the beginning to the end of the text. Based on experiments showing that human narrative comprehension seems to involve the construction of a causal network (Trabasso et al., 1989), this model makes a number of hypotheses. The associations expressed between some of the events can give the reader hints toward building a mental representation of the events. Their world and domain knowledge can enable them to sort through the expressed associations to retain and enrich the relevant ones. This can lead the reader to fill the gaps left by the text towards a causal interpretation of the events.

3.1.4 From Beginning to End

Simple generated narratives should have a structure including a clear beginning introducing a middle section and leading to a recognizable end (McKinlay et al., 2009). This model achieves this by constructing a non-hierarchical, tree-like document structure having the beginning as root and the end as the last leaf to appear in the text. The different branches link all selected events with the most important associations. This structure is called the connecting associative thread.

The type of period to be summarized should be chosen so as to have a natural beginning and end. For example, a day beginning with getting up and ending with going to bed, a working shift or a clearly delimited incident. An appropriate definition of the type of period to summarize is required before applying this model. It is possible that because of this limitation, the model would not be equally well applicable to any kind of data. In that case, some more elaborate way of determining the initial and final situations of the narrative would need to be developed.

3.2 Communicative Goal of the Generated Text

The communicative goal of the generated text in the context of this model is to communicate effectively the facts necessary to facilitate the construction of a causal network by the reader. By necessary facts, we mean the least easily predictable facts. The more easily predictable facts can, by definition, be more easily guessed by the reader. The least easily predictable facts are more difficult to guess and must be given explicitly. The sequential rule model serves to approximate how easily a fact can be predicted (given the other facts) by a reader who knows what is usual for a period a the same type of data. The least easily predictable facts are the most unusual (or least usual) of the summarized period compared to a typical period of the same kind of data. They are what makes this period unique.

The associations expressed in the generated text should give valuable hints to the reader in constructing a causal mental representation of the events. Moreover, they should generally help see the events of the period as a coherent whole if such coherence can be found. This should help the reader assimilate effectively the text's content.

The facts not mentioned in the text should be implicitly understood as "same as usual" and the reader should be able to infer them approximately from the text's content if needed. According to Niehaus and Young (2014), the reader will make such an inference if it is necessary to the comprehension of the text (criterion of necessity) and not too difficult to make (criterion of enabledness). An inference could be necessary because there is a gap in a causal chain, for example. The knowledge that the reader has of what usually happens, if the sequential association rules model that correctly enough, should enable the reader to make such inferences.

In the case of the inferences that could be triggered in the reader by the expressed associations, it is much more difficult to use the criteria of necessity and enabledness, as exactly what should be inferred or not is not known by the computer. For example, suppose an association is expressed between events A and B. If it was established that there was a causal relation between A and B, the computer could try to determine if the criteria of necessity and enabledness were fulfilled. In the affirmative, the relation could be left implicit, but if not it would need to be expressed explicitly. The problem here is that in this model, the computer does not know if there was a causal relation between A and B in the first place. There could be no causal relation or

a common cause or effect or something else altogether. In this model, the computer does not choose between these alternatives. Rather the associations are communicated and it is left to the reader to find an explanation.

3.3 Model Overview

This section presents the main components of my model of assisted temporal data interpretation using narrative generation. Figure 1 on the following page gives an overview of this model. I 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 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). The following chapters detail each of these steps.

3.4 Applications

This section sketches the type of data to which this model applies. A specific example of an application domain, Activities of Daily Living, is described in Section 3.4.1.

Genette (2007, p. 13) defines a narrative to be a discourse about an event or series of event and the various relations between them. Bal (2009, p. 5) adds that those events must be caused or experienced by actors. A data-to-text system summarizing temporal data should thus aim at generating a narrative if its users need to have a summary of what happened and how. This is to be contrasted with a descriptive summary containing only statistics about what happened. Furthermore, although this is not required by the proposed model, a generated narrative would

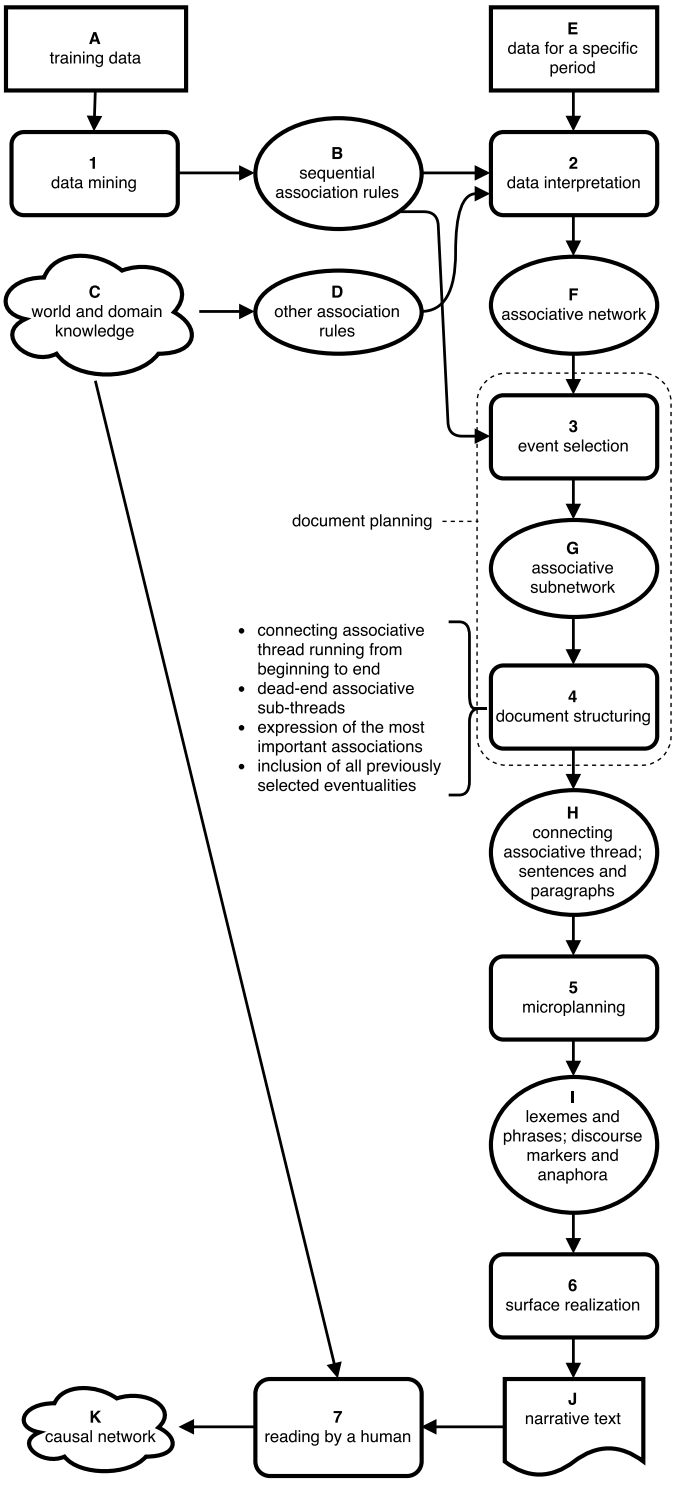

Figure 1 – Assisted temporal data interpretation model. Rectangles represent input data; rounded rectangles: computational representations; ellipses: steps; clouds: hypothesized mental representations; rectangle with S-shaped bottom side: natural language document. For ease of reference, steps are identified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

be the more relevant if the temporal data includes actions or activities. The presence of actors causing or experiencing events introduces, even if only implicitly, goals and other mental events into the causal network underlying the narrative (Trabasso et al., 1989). Note that inanimate objects or phenomena can sometimes be treated as animate for narrative purposes.

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.

Events and states can be overlapping or not. If overlapping, they can still be analyzed as a single sequence by ordering them by their start time. This does not prevent correct temporal relations to be taken into account (inclusion, overlap).

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 (Lane et al., 2010). For example, their daily use for the monitoring of the health of a person with a chronic disease or a mental disorder is currently studied (Pavel, 2013, Pentland et al., 2009). 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.) (Schneider et al., 2013). 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 (Lee et al., 2000). 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.

The following subsection describes an example of a typical application domain for the proposed model, the Activity of Daily Living domain.

3.4.1 Activities of Daily Living

Ambient assisted living technology can be used to help elderly people to live in their own house longer. Moreover, sensor equipment can be used to monitor an elderly person's Activities of Daily Living (ADLs) and detect anomalies associated with dementia early (Lalanda et al., 2010).

There are different ways of processing sensor data to detect and present possible anomalies. For example, Munstermann et al. (2012) achieve typical behavior discovery by learning a transition network from the ADL sequence data. They then use it to measure how normal a given day is and map this metric to traffic light colors.

However the normalcy of a given day is measured, health care professionals would still need to assess if there were indeed anomalies and what was their nature. For this, a more detailed access to the data is required. In my experiments, I explore a way of presenting unusual facts using NLG technology. For that sequential association rules are extracted from the event data. They are then used, together with manually written association rules, to present a textual narrative summary of a given time interval that emphasizes unusual facts. Health care professionals could then review this summary for potential anomalies with access to other sources of information. One advantage of natural language is that it can compactly express not only events but also multiple relations between events. By selecting for the generated text only the most important events and relations, the reader should not have to pore over unnecessarily detailed usual behavior.

3.5 Activity of Daily Living Data-to-text Running Example

Since this was my first experiment both with generating a narrative from extracted associations and presenting unusual facts in Activities of Daily Living (ADLs), I wanted to work on a simple yet realistic dataset. For this reason I chose the publicly available UCI ADL Binary Dataset (Ordóñez et al., 2013). This dataset was assembled to train activity classifiers that take as input raw sensor data. I do not address this task in this thesis, relying instead on the reference annotations provided as the input (but see for example the paper just cited or Fleury et al., 2010). Generating from raw data and not reference annotations would pose problems that are out of the

scope of this thesis.

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. I used only the latter in my 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 thesis. It appears in Tables I and II and figs. 2, 4, 6 and 8 to 10, on pages 26, 30, 34, 41, 45, 48 and 59, respectively.

The input for this running example consists of the data for user B as training data (A in Figure 1, p. 22) and the portion covering the day of November 24, 2012 as the data to summarize (E in Figure 1). Table I on the next page lists the 30 ADL labels of that day. On the same page, Figure 2 offers a foretaste of the sort of generated text that can be obtained with an implementation of the proposed model. The next chapters explain exactly how this is done.

In a real application and in an ideal experiment, the training data would not include the data about the day to summarize 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.

When publishing this research in Vaudry and Lapalme (2015), we have been asked about what seemed like irregularities in the UCI ADL Dataset activity labels. Sometimes the same label is repeated and one could think that it was just the same activity that continued. It has been suggested to treat such a case as a single longer activity for data-to-text purposes. However, by

Start time	End time	Activity
00:33:00	10:02:59	Sleeping
10:04:00	10:12:59	Breakfast
10:17:00	10:18:59	Toileting
10:19:00	11:13:59	Spare_Time/TV
11:16:00	11:19:59	Snack
11:30:00	11:38:59	Showering
11:39:00	11:52:59	Grooming
11:59:00	12:00:59	Grooming
12:01:00	12:02:59	Toileting
12:09:00	12:23:59	Snack
12:31:00	13:18:59	Spare_Time/TV
13:50:00	14:31:59	Spare_Time/TV
14:32:00	14:32:59	Grooming
14:36:00	15:59:59	Leaving
16:00:00	16:00:59	Toileting
16:01:00	16:01:59	Grooming
16:02:00	16:02:59	Toileting
16:03:00	16:03:59	Grooming
16:04:00	19:57:59	Spare_Time/TV
19:58:00	19:59:59	Snack
20:08:00	20:31:59	Spare_Time/TV
22:01:00	22:01:59	Toileting
22:02:00	22:16:59	Spare_Time/TV
22:17:00	22:18:59	Dinner
22:19:00	23:20:59	Spare_Time/TV
23:21:00	23:22:59	Snack
23:23:00	00:44:59	Spare_Time/TV
00:45:00	00:47:59	Grooming
00:48:00	01:48:59	Spare_Time/TV
01:50:00	09:24:59	Sleeping

Table I – The 30 ADL labels for user B on November 24, 2012.

```
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 2 – English generated text example for user B on November 24, 2012.

looking at the sensor data it can be understood why it was annotated in this way. For example, between the *Grooming* activity that finishes at 11:52 in Table I and the following activity, also *Grooming*, that begins at 11:59, the sensor data reveals that the bedroom door was used twice. In any case, I want to stress that it is out of the scope of this thesis to question the annotation process of this dataset.

3.6 Conclusion

In summary, this chapter presented a model of assisted temporal data interpretation by narrative generation. In this model, the capacities of both computer and humans are exploited to bring out a causal interpretation of the data. The computer itself only manipulates associations, formal relations of indefinite semantics which have in common that they may help the human reader in making causal hypotheses. Some association rules come from sequential data mining while others are manually derived from world and domain knowledge. The reader can follow a connecting associative thread from the beginning to the end of the generated text. This model should be most applicable to repetitive temporal data including actions or activities, such as Activities of Daily Living (ADLs).

The next two chapters describe in more detail this model and provide an example of application in the ADL domain. Chapter 4 explains how an associative network is extracted from the temporal data. Chapter 5 presents the natural language generation proper where the associative network is transformed into a narrative text. Chapter 6 discusses the evaluation of the proposed model.

CHAPTER 4

ASSOCIATIVE NETWORK EXTRACTION

This chapter explains how an associative network is extracted from temporal data according to the model introduced in Chapter 3. It is divided in two sections. Section 4.1 details how association rules are obtained. Section 4.2 presents how those association rules are used to construct an associative network for a given interval of temporal data.

4.1 Association Rules

Association rules come from two sources. Section 4.1.1 details how sequential association rules are extracted from the training temporal data using data mining techniques. Section 4.1.2 explains how world and domain knowledge can provide other kinds of association rules.

4.1.1 Sequential Association Rule Mining

For finding significant sequential association rules in the ADL data (step 1 on Figure 1, p. 22), I used the data mining techniques presented by Hamalainen and Nykanen (2008). This approach was selected because it has been successfully applied for the construction of a causal network from a video (Kwon and Lee, 2012). 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.

Generating a textual video summary using a similar technique would be an interesting endeavor. However, for that I would have first needed a reliable way of producing a sufficiently accurate textual description of an arbitrary spatio-temporal segment of a video. Generating text from ADL labels and time-stamps is easier as a first step to test the model.

4.1.1.1 Sequential Association Rule Types

In my experiments I considered a limited number of simple types of association rules in the ADL data. To select them I assumed that **temporal proximity was an indicator of potential**

causality. This is linked with the covariation principle of attribution theory (Kelley, 1973). 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.

Although temporal proximity is far from being a guarantee of causality, it is simple enough to apply as a first step. Also, the causal relation could very well be indirect or the association may instead imply a common cause between the two events. Nevertheless, it does not necessarily make the association less relevant to hint at in the generated text. The 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. I 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 (Grzegorzewski, 2011). 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 5.1.1). This assumes that the sequential associations rules are a good approximation of the knowledge that the human reader has of what usually happens. If this is correct and if necessary for his understanding of the text, the reader could guess the occurrence of the unmentioned implied event from the mentioned event (Niehaus and Young, 2014). All the corresponding probabilities are estimated by statistics, as explained in Section 4.1.1.2.

Temporal association rules are a special type of a sequential association rule. They include a categorical temporal variable such as 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 II. 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 Table I (p. 26). Rules 1, 2, and 3 are justified by the previously mentioned hypothesis 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 summarizing 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, which tends to shift the notions of morning, afternoon and evening.

Type	Sequential association rule type	Example sequential association rule 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 II – 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.

4.1.1.2 Selection Criteria

For selecting significant sequential association rules, three properties were computed for each candidate (Hamalainen and Nykanen, 2008):

— **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 part (chronological direction) or the right part (reverse chronological) of the association rule is encountered;

— **significance**: the probability of obtaining the observed counts if the events on the right part of the rule were actually independent of the events on the left part 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 B follows A or that A precedes B. This enabled me to better estimate the unusualness of each fact and thus improve content selection. This is in the case of rules referring to activities. In the case of temporal association rules, the direction of the rule means nothing in particular, 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 : one to indicate positive association rules (significantly high counts) and the other to indicate negative association rules (significantly low counts). By the latter I mean cases in which the presence of the events on the left part of the rule can be used as a predictor of the absence of the events on the right part of the rule. In other words, actual instances of these association rules are unexpected.

Frequency, confidence and significance are formalized in Figure 3 on the next page.

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.

To compute frequency, confidence and significance, I 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 were 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. I 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 ADL labels, 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 part happen with a probability around 0.1 given their left part do not give much information. They are thus less relevant.

Candidates had also 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 example rule candidates with *Toileting* given in Table II on page 30, 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 on the next page are examples of mined sequential association rules.

4.1.2 World and Domain Knowledge Rules

Causally relevant world and domain knowledge can be formalized as association rules (C and D in Figure 1, p. 22). Section 4.1.2.1 expounds on how similarities of different kinds can

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).

be a basis for useful knowledge-based association rules. Section 4.1.2.2 explains how they can also be based on specific causal knowledge.

4.1.2.1 Similarity

Association rules can be based on similarity if it is assumed that events that share some similarities may also share identical or similar causes or effects. Association rules can based on three kinds of similarity 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 in which an event occurred 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 et al. (1995): spatiality and protagonist.

The other three situation dimensions that they use in modeling 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 5.1.2.1). 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 (2014). In the context of a story with multiple actors, he associates actions having the same actor. In this way the narrative is focalized 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 the running example of this thesis.

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. An ontology contains a hierarchy of classes and sub-classes which can 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 5.1.2.1 for more details on the construction of the connecting associative thread, the narrative discourse structure proposed in this model.

Rule 6 of Figure 4 on page 34 is a simple but effective example of a manually entered association rule based on similarity of event type. It defines a *Same category* association. For the purpose of the ADL example, I arbitrarily grouped the ADL types into categories in the following manner. *Toileting*, *Grooming*, and *Showering* were placed in the category of personal hygiene activities. Those three activity types all take place in the bathroom and they share the same general function, to take care of the body. *Breakfast*, *Lunch*, *Dinner*, and *Snack* were

grouped as eating activities. Those four activity types involve food and the use of the kitchen. *Spare_Time/TV*, *Leaving*, and *Sleeping* were kept in separate categories, because they were considered to have significantly different functions from the others and each other. In addition, they are performed in different locations: the living room, the entrance/outside, and the bedroom, respectively.

In the example ADL sequence of Table I (p. 26), 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 II (p. 30). 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 obtaining a more appropriate connecting associative thread in document planning.

4.1.2.2 Specific Causal Pattern

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 (2016). They encode some of the world knowledge necessary to interpret the behavior of agents in the context of the Triangle Choice of Plausible Alternatives (Triangle-COPA) challenge problems (Maslan et al., 2015). Triangle-COPA is a set of 100 short sequences of events involving geometrical shapes having humanlike behavior. Each is followed by a question about the interpretation to give to this sequence of events. A choice of two plausible interpretations is given. The goal is to select the same one that a human would choose. The correct answer was determined by consensus among multiple human raters. Both the sequence of events and the possible interpretations are given in both English and first order logic.

Figure 5 on the next page shows some examples of commonsense axioms presented by Gordon (2016). They are here translated in a notation similar to the one used for sequential

association rules. All those association rules present plausible causes for an agent x to chase an agent y. They also include an estimation of the likelihood that the antecedent (here the left part of the rule) implies the consequent (here the right part). Rule 1 states that if x is playing with y, x may chase y with a likelihood of 0.2. Similarly Rule 2 gives the likelihood of the chasing behavior if x was angry at y. Rules 3 and 4 propose that the chase may be caused by x's goal to rob y or make y afraid of them.

As can be seen, those examples include a manual estimation of the probability associated with each probabilistic implication. That could make it possible to use this kind of world and domain knowledge rule in the same way as mined sequential association rules in document planning. The likelihood attached to each rule could be used in the same way as the confidence of sequential association rules to determine which events should be selected, if this equivalence is accepted. See Section 5.1.1 for more details on how confidence is used in event selection.

Another example of manually written rules encoding specific causal patterns is the error explanation specifications of Farrell et al. (2015). In this work, regular expressions are used to define explanation specifications for error trace data. In this way the relation between a malfunction and its possible causes can be identified. Regular expressions could enable the capture of complex causal patterns by association rules as part of this model.

1. $playWith(e1,x,y) \xrightarrow{0.2} chase(e,x,y)$
2. $angryAt(e1,x,y) \xrightarrow{0.2} chase(e,x,y)$
3. $goal(e1,e2,x) \wedge rob(e2,x,y) \xrightarrow{0.3} chase(e,x,y)$
4. $goal(e1,e2,x) \wedge afraid(e2,x,y) \xrightarrow{0.5} chase(e,x,y)$

Figure 5 – Specific causal pattern-based association rules from the Triangle-COPA domain adapted from Gordon (2016). e, $e1$, and $e2$ are variables representing events. x and y are variables representing humanlike agents. Each arrow represents a probabilistic implication. The number over each arrow is the likelihood that the consequent is implied by the antecedent.

4.2 Data Interpretation

Data interpretation (step 2 of Figure 1, p. 22) consists of searching the data to summarize for instances where an association rule applies. Sequential associations are derived from rules such as Rules 2 to 5 from Figure 4 (p. 34). They are shown as arrows going from one row to another at the left of Figure 6 on page 41. 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 6. *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.

Section 4.2.1 discusses what role should play the chronological direction of the sequential association rules in interpreting the data. Section 4.2.2 introduces associations that are derived from other associations after all instances of association rules have been identified.

4.2.1 Rule Direction and Causal Interpretation

As explained in Section 3.1.3.1, when I decided to try mining reverse chronological sequential association rules, my 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 6 on page 41. 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 premeditate some preparation for it.

However, the underlying goal hypothesis does not seem to hold in all cases. Again by looking at Figure 6, 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, than 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 does not contain enough information about the events on which the premeditation is based. In short, how much information does 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 I claim to identify causal relations? This is one reason why I 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.

4.2.2 Derived Associations

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 6 on the next page.

4.3 Conclusion

In this chapter were presented the methods used for constructing an associative network from temporal data. First sequential association rules are mined from training data. They are selected by computing their confidence and significance based on statistics. World and domain knowledge rules can also be manually formulated. They can be based either on some similarity between events or on some specific causal pattern difficult to discover statistically.

When the input period to summarize has been chosen, its temporal data can be interpreted with the help of the association rules. Occurrences where an association rule applies are identified and the relevant events are associated. Then some extra associations are derived.

The next chapter presents how the extracted associative network serves as a basis for the generation of a narrative text. It covers the NLG steps of document planning, microplanning, and surface realization. Finally, it discusses the interpretation of the text as a narrative by the human reader.

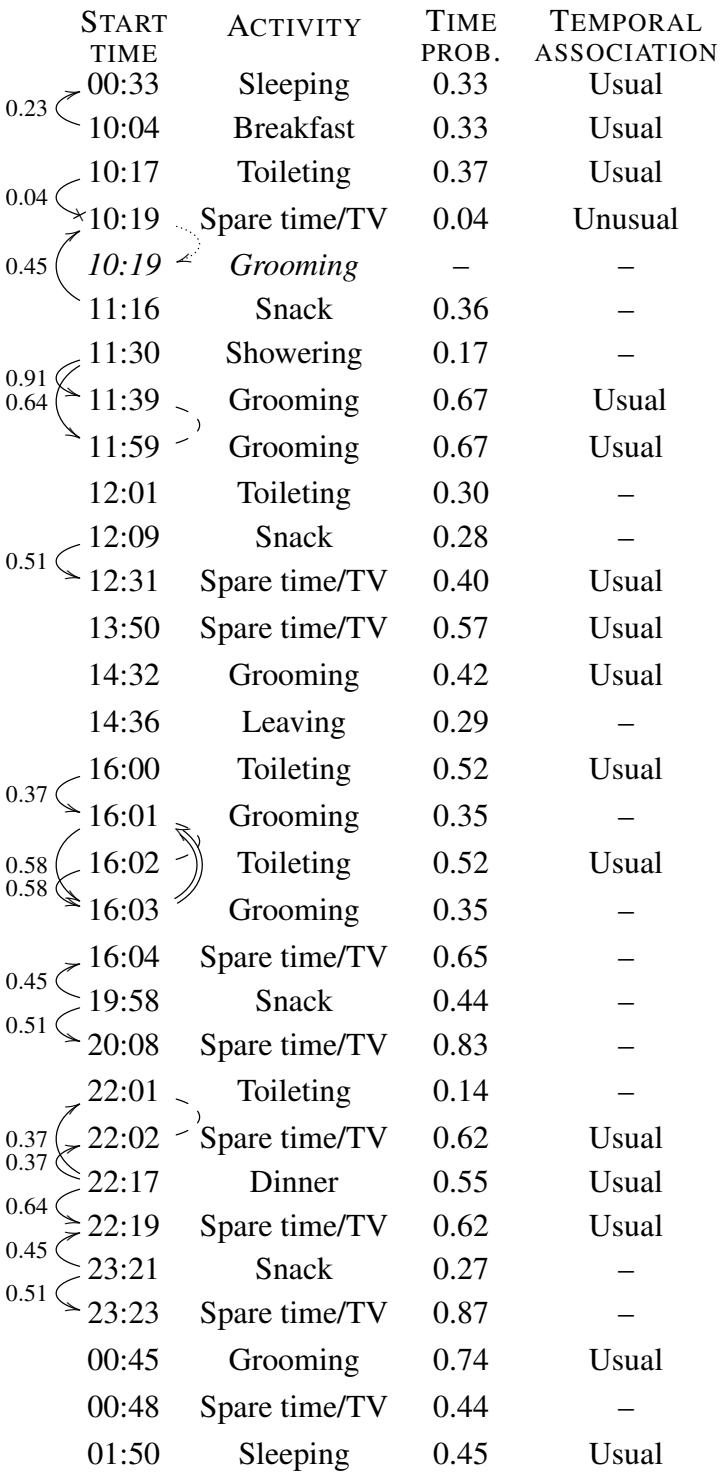

Figure 6 – Associative network for user B on November 24, 2012. Sequential associations are on the left. The X-headed arrow represents an unexpected association. On the right are *Instead* (dotted), *Conjunction* (dashed), and *Repetition* (double). Under *Time prob.* is the confidence of the corresponding temporal association rule candidate. If the latter was selected as an expected or unexpected association rule, it is marked as *Usual* or *Unusual*, respectively.

CHAPTER 5

NARRATIVE GENERATION

After extracting an associative network from temporal data, it can now be used to plan the content and structure of the narrative and generate natural language text. This chapter details the workings in the proposed model of the three stages of the standard NLG pipeline: document planning, microplanning, and surface realization (Reiter and Dale, 2000). Section 5.1 expounds on the document planning stage, in which content is selected and structured. Planning at the level of the sentence and below is described in Section 5.2 on microplanning. Details about surface realization can be found in Section 5.3. Finally, the hypotheses that the model makes about how the generated text is interpreted as a narrative by the human reader are discussed in Section 5.4.

5.1 Document Planning

Document planning is the most important section of this chapter. The associative network given by the data interpretation stage serves as raw material in this task. First events are selected by taking into account the confidence computed for each sequential association present in the network. This is described in Section 5.1.1. Then document structuring can take place, where the connecting associative thread is determined and the text plan is segmented into paragraphs and sentences. This is the object of Section 5.1.2.

5.1.1 Event Selection

As can be seen on Figure 1 (p. 22), event selection (step 3) takes as input the associative network and outputs a sub-network of its input. The purpose of this step is to select the events that are the most unusual, that is the less probable according to the sequential rules model. This unusualness depends on associations for associated events and on time of occurrence for isolated events. This step is called event selection and not simply content selection because final association selection takes place later, during document structuring. Only associations that are

used to build the document structure are ultimately retained in the document content. However, document structuring can only select associations between events selected in the event selection step.

Event selection has one parameter: a maximum probability threshold. The purpose of this threshold is to generate summaries that can vary in length depending of 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 7 on the following page.

The purpose of the first part of the algorithm 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 second part of the algorithm makes sure that associations that justified the selection of an event in the first part also have their other argument selected. This way those associations can be included in the document structure.

The last condition of the algorithm targets events that are not implied by any other event or time. They 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. Events that are not the second argument of any selected sequential association rule were at first selected using their prior probability of happening in any time step, regardless of time of occurrence. A problem with those events is that they have none or few sequential associations linking them to other selected events. This can lead to texts with less or less diverse rhetorical relations marked by discourse markers. Yet some of those events that have low prior probability prove to be easier to predict when the probability conditioned on time is used for determining their unusualness. According to my experiments with ADL data, this strategy often eliminates additional recurrent activities that are isolated from the other activities in terms of sequential associations.

```
procedure SELECT(event e)          ▷ returns true if e is to be selected and false otherwise
    if ∃(x --p--> e) ∈ associations then                          ▷ x is an event or a time
        for all x --p--> e ∈ associations do
            if p ≤ threshold then
                return true
            end if
        end for
        return false
    else                          ▷ y is an event, P(e|t_e) is the probability conditioned on time of e
        return (∃(e --p--> y) ∈ associations, p ≤ threshold) ∨ (P(e|t_e) ≤ thresold)
    end if
end procedure
```

Figure 7 – Event selection algorithm.

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 the case of the running example presented in this thesis, 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 summaries while still displaying interesting textual phenomena. Figure 8 on the next page 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 remainder of the running example is 0.3.

If they have not already been selected, the first and last events of the period to summarize will be added to the selected events to become the initial and final situations of the connecting associative thread. This is fully explained in Section 5.1.2.1.

5.1.2 Document Structuring

Document structuring (step 4 in Figure 1 on page 22) takes as input the associative sub-network that includes only the most unusual events, as provided by the previous step of event

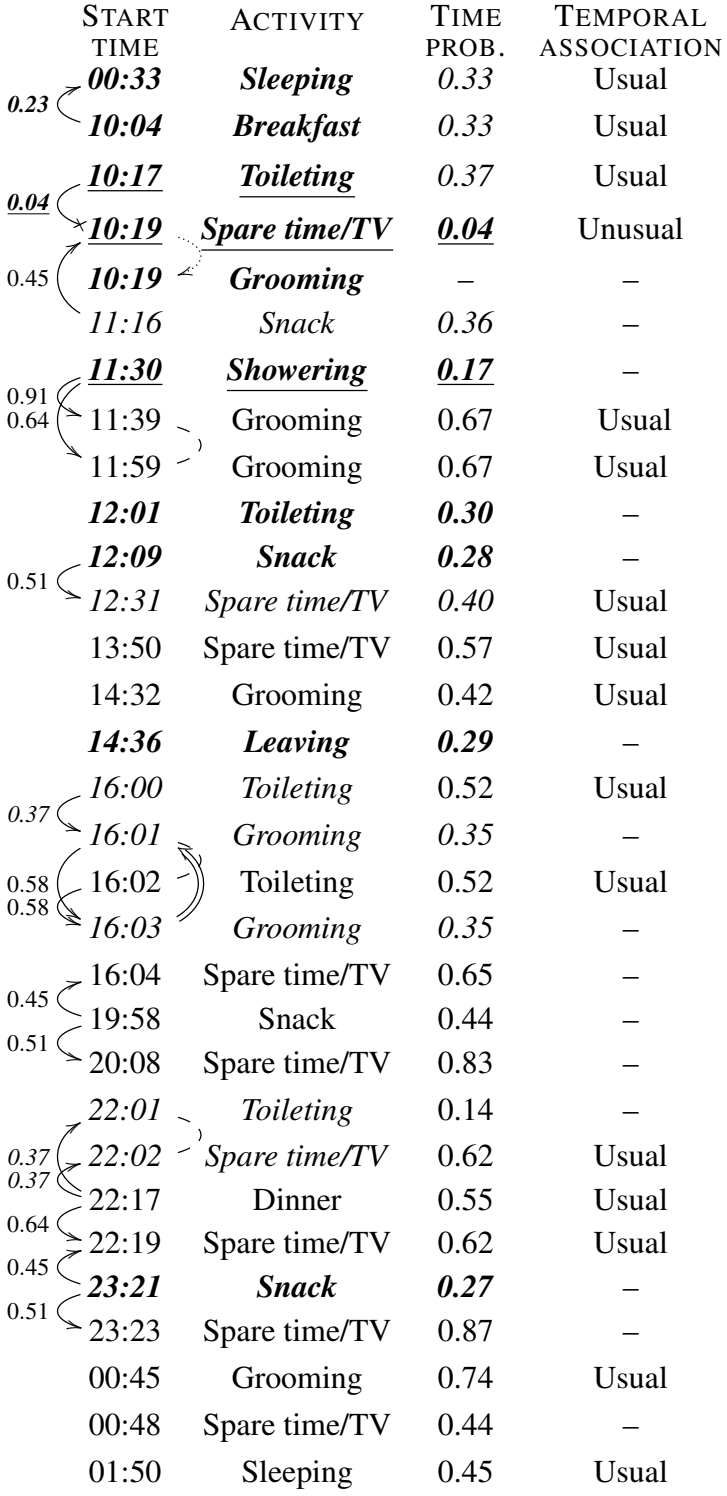

Figure 8 – Associative network for user B on November 24, 2012. Selected events are shown for maximum probability 0.2 (underlined), 0.3 (bold), and 0.4 (italics). Sequential associations are on the left. The X-headed arrow represents an unexpected association. On the right are *Instead* (dotted), *Conjunction* (dashed), and *Repetition* (double).

selection. Its output must be 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. The most important part of document structuring is the determination of the connecting associative thread, described in Section 5.1.2.1. Section 5.1.2.2 explains the next step, where the document plan is divided into paragraphs, sentences, and phrases. Lastly, Section 5.1.2.3 presents the mapping that has to be made between the associations of the associative network and the rhetorical relations that will be marked in the text.

5.1.2.1 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 was highlighted by a comparison with human written texts (McKinlay et al., 2009). The connecting associative thread, as its name suggest, must also connect all the previously selected events with appropriate associations, so that they form as much as possible a coherent whole.

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 8 on page 45). 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 8). 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. Table III on the following page gives the association preference values used in the example.

The association preference is combined (by averaging) with the relative temporal distance in order to favor 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

Association type	Association preference
Expected sequence	0.40
Unexpected sequence	0.35
Instead	0.05
Conjunction (with p the association preference of the association that the coordinates have in common)	$0.67 \times p$
Repetition	0.40
Same category	0.60
Temporal proximity (default)	1.00

Table III – Association preferences for each association type. 1.0 means as far as possible and 0.0 mean as close as possible. The actual adjacency preference for conjunction is a coefficient applied to the adjacency preference of the association that the coordinates have in common.

of the period to summarize. Averaging is used to combine the two because it preserves the range of values (contrarily to a simple sum) and is linear (contrarily to multiplication, for example). The linearity makes it easier to understand the impact of increasing or decreasing a parameter. A weighted average could be used instead to adjust the influence of chronological order on textual order. Favoring temporal distance would tend to associate temporally consecutive events.

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 toward the final situation. The connecting associative thread connects every event through the main thread and the dead-end threads. This is illustrated in Figure 9 on the next page.

As indicated in Table III, the association preference for the *Conjunction* association works differently from the others. It is actually a coefficient applied to the adjacency preference of the associations that each have the coordinates as one of the arguments and share the other argument. This usually has the effect that a *Conjunction* association will be selected just before

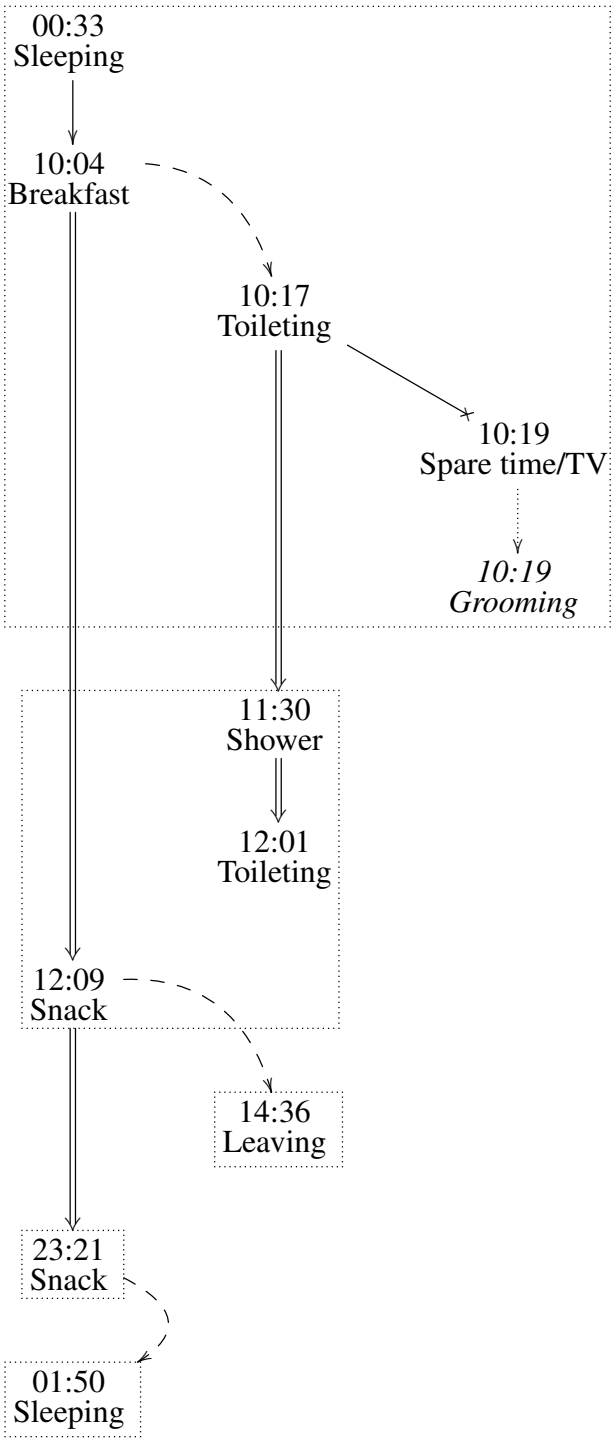

Figure 9 – Connecting associative thread for user B on November 24, 2012. Arrows represent associations: simple: expected sequence; X-headed: unexpected sequence; double: *Same category*; dotted: *Instead*; curved and dashed: temporal proximity. Paragraphs are boxed. The vertical order of presentation is the order of mention in the generated text (Figure 10 on page 59). For event selection, the maximum probability threshold was set to 0.3.

this set of common associations. Because a tree contains by definition no cycle, only one of those associations will be selected as part of the connecting associative thread. Also, the coordinates will always be grouped closer in the segmentation step (Section 5.1.2.2) than with the shared argument of the common associations.

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.

5.1.2.1.1 Further Narrative Structure Comments Research on causality in narrative comprehension has uncovered that events on the causal chain going from the beginning to the end of the story are more often recalled than those on dead-end parts of the causal network (Trabasso and van Den Broek, 1985). In the future, it may be interesting to verify if events on the associative sub-threads are less remembered than those on the main associative thread. If this is the case, the content structuring algorithm should be modified to optimize the importance of the expressed associations together with the proportion and importance of the events included in the main associative thread. However, taking into account the relative temporal distance in the computation of the minimum spanning tree already tends to avoid a too short main associative thread.

In the ADL dataset used for the running example, the system has just one state variable: the activity type of the user. Consequently, it suffices that the initial and final situations simply report on the beginning and end activities. In a more complex situation, more than one variable may evolve over time. As noted in McKinlay et al. (2009), generally the description of the initial situation should include information just on the variables that will be relevant in the middle section of the narrative and the final situation should provide a concluding statement on the state of those variables at the end of the period. This aspect of narrative planning is not further developed as part of this thesis.

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 period to summarize. 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. A separate segmentation algorithm was used to divide the dataset into such cycles before applying the data-to-text model. The heuristic was to separate days at *Sleeping* activities lasting more than 1 hour and starting at least 16 hours after the start of the last separating activity. It is not guaranteed that a natural temporal division of the data will always be possible. In case the period to summarize does not begin and end with naturally good initial and final situations, it may be necessary to develop more elaborate ways to determine them.

The connecting associative thread has a tree structure, but it is a tree based on associative dependencies between events, not a constituency tree based on relations between recursive text segments, like the rhetorical structure of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In my first experiments, I tried to use the latter kind of tree as the document structure for generating narratives from sequential associations. One of the major problems was the difficulty to find relations between text segments placed higher up in the hierarchy. Those would need to be more abstract relations that the relations found between single events, because they involved complex configurations of event descriptions. I was not able to produce a hierarchical structure fully labeled with rhetorical relations with the bottom-up method I was employing. Also, it was difficult to manage temporal ordering properly using such a hierarchical structure. The connecting associative thread algorithm has the advantage of producing structures fully labeled with associations. This is a prerequisite for the integration of the greatest number of planned rhetorical relations in microplanning.

5.1.2.2 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 (Terluin, 2008, p. 83). 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 Section 5.1.2.1. 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. Paragraphs are boxed in Figure 9 on page 48.

Between sentence breaks, consecutive event-describing clauses are grouped by two recursively 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 be used in microplanning to determine syntactic relations between phrases.

In a previous version of the model, paragraph and sentence segmentation depended solely on the distance score computed previously. The association preference was also an aggregation preference. It determined if the arguments of the association preferred to be put in the same sentence, in the same paragraph, or in separate paragraphs. The association preferences thus played a double role: indicate in what order association types should be preferred and indirectly specify the number and locations of paragraph and sentence breaks. By creating dedicated parameters for segmentation, structuring and segmentation were disentangled and the model made more transparent and flexible. The association preferences still influence paragraph and sentence segmentation by ordering candidate break points, but not longer determine the number of paragraphs and sentence breaks. One consequence is that a text in which there are very few sequential associations will not necessarily result in a text with the content split into many

very small paragraphs, as was the case before. What would have been the content of several smaller paragraphs will be grouped in a bigger paragraph to keep the average number of events introduced in each paragraph roughly constant.

5.1.2.3 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. (2012). The many-to-many mapping means that how associations are detected is in principle independent of how they are expressed.

Although the document structure is not a rhetorical structure in the sense of RST (Mann and Thompson, 1988), the concept of rhetorical relation I 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 IV on the next page 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.

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 IV – 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.

5.2 Microplanning

Microplanning (step 5 of Figure 1, p. 22) translates the document structure into a lexico-syntactic specification ready for surface realization. This section addresses microplanning at three levels: the description of a single event, combining event descriptions inside a sentence, and joining sentences to form the paragraphs of a text.

As part of my experiments with ADL data, texts can be generated in English or French on demand. This did not affect document planning, but it did impact microplanning and surface realization. Accordingly, this section includes where appropriate considerations on bilingual generation for relatively close languages such as English and French. Note that for less closely related languages, bigger adjustments may need to be made, possibly including changes in document planning.

5.2.1 Description of a Single Event

There are two scenarios in which a description of an event is needed. The first one is when the event is first mentioned. In that case the description takes the form of a clause. The second one is when the event description plays the role of an anaphor. Here the event is described by a noun phrase or a verb phrase that functions as a noun phrase. The latter can be a gerund, in English, or an infinitive proposition, in French.

Depending on the event type and the language, the type of an event can be lexicalized by a

single lexical item, a collocation, or a fixed expression. An event's argument can be lexicalized by a noun phrase or replaced by an appropriate pronoun when required by the context. It can then play the role of subject, object or complement of the phrase representing the event type. Other event properties can be lexicalized in the form of adjectival or adverbial modifiers.

All of those are hand-coded in the form of lexico-syntactic templates. Those are far more flexible than regular templates, because they are described in terms of lexical items and syntactic functions, not (only) fixed strings. They can thus be adapted to their final syntactical environment by being subject to inflection and the addition of further modifiers, among others.

Table V on the following page gives examples of lexico-syntactic templates for the description of two activity types: *Grooming* and *Breakfast*. The lexico-syntactic templates are given for both the first mention and anaphor cases mentioned earlier and also for both English and French. The *Grooming* template is specified using only one lexical item, while others include a collocation (for example, *eat his/her breakfast*) or a fixed expression (French *faire sa toilette*). Using an appropriate surface realizer, it is easy to modify the structures given by those lexico-syntactic templates to get a version with a different tense or add a modifier. For example, one can add the *perfect* feature and a prepositional complement expressing time to the *Grooming* anaphor template to obtain: *having groomed at 10:00 PM*. Moreover, the templates let the surface realizer take care of agreement between, for example, the French possessive determiner *son* and the name it specifies: *son petit-déjeuner* (masculine) but *sa toilette* (feminine).

To decide when to add an absolute time phrase such as *at 10:00 PM*, information on rhetorical relations is needed. If the event is in a Sequence rhetorical relation with its parent in the connecting associative thread, a relative time marker (*then*) will be added and an absolute time phrase is presumably not needed. When it is present, the absolute time phrase will be realized in front of the clause unless this position is occupied by a subordinated clause. In that case it will be placed at the end of the clause. An event duration phrase is also added to descriptions of activities with a typically long but variable duration: *Sleeping*, *Leaving*, and *Spare time/TV*. This temporal modifier strategy has some similarities with that of Hunter et al. (2012), which also focuses on long events and events whose times are not expressed by reference to other events.

The examples of Table V illustrate that English and French morphology and syntax differ in

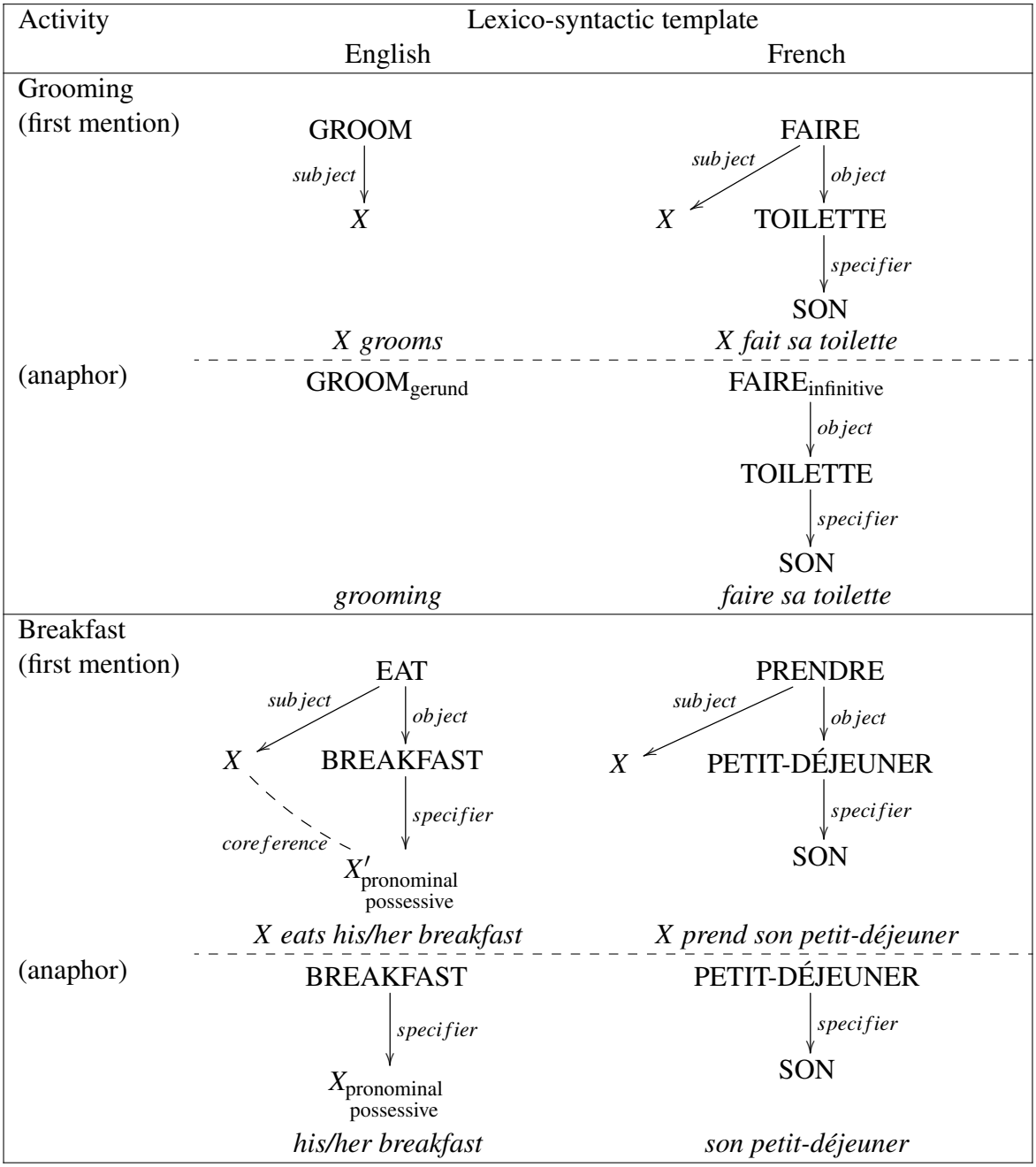

Table V – Example English and French event description lexico-syntactic templates used in the ADL example. Below each is a natural language realization in the present tense.

their treatment of possessive. In English, the possessive determiner must be treated as a pronoun in that it must reflect the biological gender and animacy of the noun phrase it corefers with. By contrast, in French, the possessive determiner, like any French determiner, must agree in grammatical gender and number with the noun it specifies. The type of information needed to handle the morphology of possessives in this language pair is thus not the same. However, although it must be taken into account when designing the lexico-syntactic templates, it is not a difficult problem to overcome given the availability of a surface realizer that handles this correctly.

5.2.2 Combining Event Descriptions to Form a Sentence

To combine event descriptions to form the specification of a sentence, the microplanning stage relies on the sentence plans assembled during document planning and described in Section 5.1.2.2. Each of these binary trees is traversed depth first. When a leaf is visited, a specification of the corresponding event's description is produced from the lexico-syntactic templates described in the previous subsection. When an internal node is visited, the rhetorical relations linking the two children nodes are expressed with appropriate discourse markers. The intrasentential discourse markers used in the ADL example for each rhetorical relation are given in Table VI. Those markers are then used to assemble the lexico-syntactic specifications obtained from the children nodes. Specifically, the marker is attached to the last mentioned argument of multinuclear relations and to the satellite of nucleus-satellite relations. Depending on the marker, clauses may be subordinated or coordinated and other phrases added as complements.

Rhetorical relation	Discourse marker	
	English	French
Sequence	then	ensuite
Contrast	but	mais
Instead	instead of	au lieu de
Conjunction	and	et
Repetition	again	encore

Table VI – English and French intrasentential discourse markers used in the ADL example for each rhetorical relation.

5.2.3 Combining Sentences to Form the Paragraphs of a Text

The marking of rhetorical relations between sentences is handled differently than between clauses of a single sentence. 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.

Theories of intersentential coreference such as Centering Theory (Grosz et al., 1995) 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. 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 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 (Terluin, 2008, pp. 68–69). 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 9 (p. 48) is *Shower 11:40*. Since 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 10a on the following page). 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*. The intersentential discourse markers used in the ADL example for each rhetorical relation are given in Table VII.

Rhetorical relation	Discourse marker(s)			
	English		French	
	alone	with anaphor	alone	with anaphor
Sequence	then	after	ensuite	après
Contrast	however		toutefois	
Instead	(negation)		(negation)	
Conjunction	in addition additionally also	in addition to beside aside from	de plus également aussi	en plus de outre
Repetition	again		encore une fois	

Table VII – English and French intersentential discourse markers used in the ADL example for each rhetorical relation.

5.3 Surface Realization

Surface realization (step 6 of Figure 1, p. 22) was performed using the SimpleNLG-EnFr Java library (Vaudry and Lapalme, 2013). During surface realization, the syntactic and lexical specifications are combined with the output language grammar and lexicon to generate formatted natural language text. As explained in Section 5.2, a version of the lexico-syntactic templates used in microplanning for the ADL experiments was written for each of the two output languages, English and French. In combination with SimpleNLG-EnFr, this enabled bilingual generation.

English and French example generated texts corresponding to the preceding figures are given in Figure 10 on the following page.

```
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.
```

(a) English
(reproduced from Figure 2)

```
OrdonezB samedi, 24 novembre
2012 0h33 - dimanche, 25
novembre 2012 9h24
------------------------------

OrdonezB s'est levé à 10h02
et ensuite, il a pris son
petit-déjeuner.  Comme
d'habitude à 10h17, il est
allé à la toilette, mais
ensuite, il a passé 1 heure
au salon de façon inattendue
au lieu de faire sa toilette.

En plus d'être allé à la
toilette à 10h17, il a
pris une douche à 11h30.
Également, à 12h01, il est
allé à la toilette.  Outre son
petit-déjeuner de 10h04, il a
pris une collation à 12h09.

À 14h36, il est parti pendant
1 heure.

En plus de sa collation de
12h09, il a pris une collation
à 23h21.

Comme d'habitude à 1h50, il
est allé se coucher.
```

(b) French

Figure 10 – English and French generated text examples for user B on November 24, 2012. The maximum probability threshold was set to 0.3.

5.4 Human Reading

Finally, in step 7 of Figure 1 (p. 22) 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.

I 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 of that share some similarity sometimes have the same cause or effect or the same type of cause or effect. Other association rules based on specific causal patterns could also give useful hints. See Section 4.1 for more details about how the different types of associations may help the reader make causal hypotheses. In any case, the reader can choose to ignore irrelevant associations.

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 of his knowledge. On 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 happening.

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.

5.5 Conclusion

In this chapter, I described an NLG pipeline for generating a text interpretable as a narrative from an associative network. The most important stage is document planning. In that stage, events are selected based on the confidence of the sequential associations they are connected

with. The goal is to keep only the most unusual events and leave implicit the events that can be inferred from the others according to the sequential association rule model. Next a connecting associative thread is determined that structures the text by connecting all selected events and comprising a thread leading from the initial situation to the final situation. Then the document plan is segmented into paragraphs, sentences, and phrases. This operation is parameterized by a target average number of sentences per paragraphs and a target average number of events introduced by sentence. After mapping associations to rhetorical relations, comes the following stage: microplanning.

Microplanning takes the document plan and, using lexico-syntactic templates, assembles a specification that can serve as input to a surface realizer. The lexico-syntactic templates define how an event in described in terms of lexical items and syntactic structures. The event description specification can then be assembled using discourse markers based on the rhetorical relations to be expressed. When combining sentences, the main event and preceding main event are computed for each sentence. This serves to determine if an anaphor is needed to help coreference resolution by the reader. In the ADL experiments, a version of all lexico-syntactic templates was hand-coded for each of the two output languages: English and French.

The lexico-syntactic specification for the whole text is realized as natural language by the SimpleNLG-EnFr surface realization Java library. This library enables bilingual generation in both English and French in conjunction with the bilingual specifications provided by microplanning.

Finally, the model assumes that the human reader interprets the generated text as a narrative and constructs a causal mental representation of the text's events.

In the next chapter, I will discuss the evaluation of the assisted interpretation model and of the generated texts. An evaluation of the the textual quality of the generated texts has been conducted and its results will be presented.

CHAPTER 6

EVALUATION

Evaluations of NLG systems have been classified as either intrinsic or extrinsic (Belz and Reiter, 2006). Intrinsic methods involve testing the properties of the form the NLG output takes, the generated texts in and of themselves. Extrinsic evaluations look at the impact of the generated texts on task performance and generally at how well the NLG system accomplishes its overall function. The content of this chapter is divided according to this dichotomy. Section 6.1 describes an intrinsic evaluation of the text quality of generated ADL reports by human judges. Its method and results are detailed. Section 6.2 discusses possible extrinsic evaluation designs for testing communicative goal attainment of the texts generated with the proposed model.

6.1 Intrinsic Evaluation of Text Quality by Non-expert Judges

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. Section 6.1.1 details how the reports were generated, the participants recruited, and the evaluation forms designed. Section 6.1.2 discusses what the results show for each evaluation criterion. Section 6.1.3 compares the presented evaluation with previous temporal data-to-text evaluations.

All the generated texts, evaluation forms, and judge answers of this evaluation are publicly available on the web [1].

6.1.1 Method

6.1.1.1 Data-to-text report generation

For the text quality evaluation, as for the running example, texts were generated using data from the UCI ADL Binary Dataset (Ordóñez et al., 2013) as input for training and generation. See Section 3.5 for more detail on this dataset. Reports were generated from both user A and B data. As discussed in Section 3.5, because of the small size of the dataset, all the data for a

1. http://www-etud.iro.umontreal.ca/~vaudrypl/ADL/eval/

given user was used as training data for that user's reports. To assemble the evaluation corpus, a report was generated for the 32 complete days of the dataset. As discussed in Section 5.1.2.1.1, a day was defined heuristically as starting and ending with a *Sleeping* activity lasting more than 1 hour and starting at least 16 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 10a on page 59.

6.1.1.2 No Baseline or Gold Standard

In a scenario where the reports to be generated already exist in one hand-authored form or another 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 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 and I lacked the time and resources to recruit them as part of my thesis research.

6.1.1.3 Evaluation Criteria

The design of the evaluation criteria for this experiment was mainly based on the evaluations presented in Callaway and Lester (2002, p. 241) and Hunter et al. (2012, p. 168). Both asked human judges to rate generated narrative texts. The former used the following criteria for the evaluation of their fictional narrative generation system: Overall, Style, Grammaticality, Flow, Diction, Readability, Logicality, Detail, and Believability. The latter asked users to rate Understandability, Accuracy, and Helpfulness of the generated text in real use. Those criteria were considered for application to the ADL reports evaluation. Readability was judged to be too vague and encompassing. Level of Detail, Accuracy and Helpfulness were not applicable outside of a realistic context and with non-expert judges. Believability is not applicable to non-fictional texts. Logicality was discarded in favor of Understandability, which seemed easier to understand. Diction was renamed Vocabulary for clarity. Overall, Style, Grammaticality, Flow seemed appropriate and were kept. Below are the six chosen criteria and the corresponding questions as they appeared on the evaluation form.

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 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.

6.1.1.4 Evaluation Form

A sample of a blank evaluation form can be found in Appendix I. Since each form repeated the same questions for each of four or five texts, only the first three pages of this form, covering

the questions for the first text, are included in the appendix. This appendix contains a paper version of the form divided into Letter sized pages, but the real form was viewed in the form of a Google Forms web page by the judges. The judges could view and answer all the questions for one text on the same page. They could then continue to the next text evaluation by clicking a button at the bottom of the page. As can be seen in the appendix, the evaluation forms were presented in English only.

6.1.1.5 Participants

13 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 persons who at least approached native ability in English were accepted as judges.

6.1.1.6 Assignment of Evaluation Forms

13 judges evaluated four to five generated texts each, so that 28 texts were evaluated by two judges each and 4 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. Table VIII details the distribution of texts among forms and the number of judges that received and filled each form.

Form #	Texts	Number of judges
1	A1, A9, B3, B11, B19	2
2	B20, B12, B4, A10, A2	2
3	A3, A11, B5, B13	2
4	A4, A12, B6, B14	1
5	B15, B7, B2, A13, A5	2
6	A6, B8, B16, B18	2
7	B17, B10, B9, A8, A7	2

Table VIII – Texts and number of judges for each evaluation form. The text numbers refer to the user (A or B) and the day number in each user sub-dataset. Each judge filled only one form.

6.1.2 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 11 on the next page.

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 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.

6.1.3 Comparison with Previous Temporal Data-to-text Evaluations

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

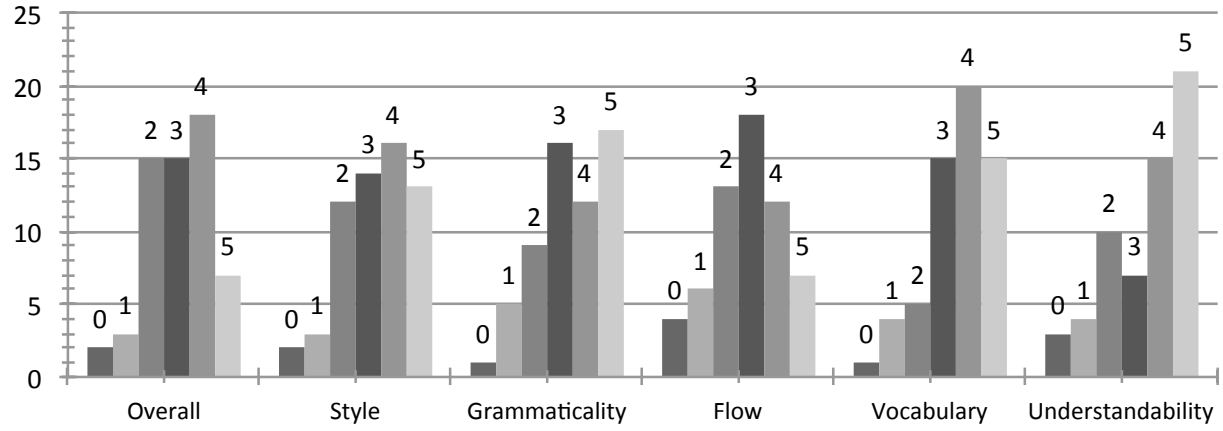

Figure 11 – Results of the text quality evaluation. On the vertical axis is the number of evaluations and on the horizontal one are the ratings for each criterion.

tems.

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 (Hallett, 2008). Some used human ratings, similarly to the evaluation presented in this thesis, to evaluate text quality. Wanner et al. (2010) evaluated in this way 16 air quality bulletins generated in each of 5 languages. The results indicate that Comprehensibility, Fluency, Content relevance, and Level of detail needed improvement. Bouayad-Agha et al. (2012) had 51 generated football match summaries rated and reported good results for both Intelligibility/Grammaticality and Fluidity. Note that the number of rated generated texts in those studies and this thesis are of the same order.

Since football match summaries are regularly written by journalists and the data behind them is available, Bouayad-Agha et al. (2012) also compared content selection between human and generated summaries for 36 matches. They concluded that recall was good, but precision needed improvement. In this thesis, 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 latest is the discourse analysis carried out in the Babytalk BT-45 project on 3 generated texts and 3 human written ones (McKinlay et al., 2009). As mentioned in Section 2.5.2, this analysis uncovered problems

in narrative structure and in expressing temporal information. For the same project, an off-ward decision experiment was also conducted (Portet et al., 2009). In this experiment, medical experts had to determine the required action based on historical data on 24 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. In an on-ward evaluation as part of the BT-Nurse project, 148 generated texts were rated by nurses (Hunter et al., 2012). 90% of ratings stated that the text was understandable, 70% accurate, and 59% useful. Lastly, 3 experts and 2 members of the target audience rated 3 versions of a text for the Tag2Blog project (Ponnamperuma et al., 2013). One first generated version contained only indications of 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, than the second versions. Compared with the evaluation presented in this thesis, evaluations with experts require much more resources. Considering that such resources were not available as part of this thesis 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. (2010) and Bouayad-Agha et al. (2012), 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 to show that the proposed approach is interesting and worthy of future research.

6.2 Extrinsic Evaluation Designs

The intrinsic evaluation described in the preceding section has given an idea of the quality of the generated texts. To test the validity of the model as one in which generated texts assist a human reader in forming a causal representation of the events, however, an extrinsic evaluation is

needed. That causal relations play a role in human-written narrative comprehension has already been verified experimentally (Trabasso and van Den Broek, 1985). However, would a generated narrative produced with only indirect and incomplete evidence and background knowledge about causal relations trigger the same phenomena in the reader? This still needs to be tested. Another question is whether the mental representation derived from such a generated narrative would constitute a correct and useful enough interpretation of the data. To give ways of answering those questions, this section sketches two extrinsic evaluation designs: a memorization experiment (Section 6.2.1) and a domain-specific task-based evaluation (Section 6.2.2).

6.2.1 Evaluation of Unusual Event Recall

Part of the text's communicative goal, as stated in Section 3.2, is to make the reader assimilate the facts that are characteristic of the summarized period. To test this I imagined an experiment in which the subjects have to memorize a series of five consecutive days. The data for each day is first presented either in the form of a raw data table, a visual rendering of the association sub-network, or a generated text. (The second form serves to isolate the effect of the text itself.) The same days are then presented again in random order to test reordering performance. In this phase, the data is presented in the form of an approximate chronology with icons representing activities. One example of this last presentation format is given in Figure 12 on the following page. Its purpose was to avoid reproducing in the reordering phase the visual characteristics of the presentation formats used for memorization.

The hypothesis is that if the generated texts are better at communicating the most unusual facts, they will make it easier to recognize which day is which. Therefore, the subjects will make fewer mistakes in the reordering task. One advantage of this experimental design would be that it is domain independent. Additionally, a gold standard for what constitutes an unusual fact is not needed.

By doing a dry run of this experiment, I found out that the subject could easily find during the memorization step trivial information that could be used to differentiate days. For example, they could memorize bedtime or the number of *Spare time/TV* activities. Therefore this design could not be used as is to verify the research hypothesis. One modification that could hinder

Start time	Activities
8 AM	
9 AM	
10 AM	
11 AM	
12 PM	
1 PM	
2 PM	
3 PM	
4 PM	
5 PM	
6 PM	
7 PM	
8 PM	
9 PM	
10 PM	
11 PM	
12 AM	
1 AM	

Activity icon	Activity name
	Getting Up
	Sleeping
	Toileting
	Grooming
	Showering
	Spare Time/TV
	Leaving
	Breakfast
	Lunch
	Snack
	Dinner

Figure 12 – Chronology and icon legend used in the memorization experiment dry run. This presentation format was used to test recall in the second phase of the experiment. Its purpose was to avoid reproducing visual characteristics of the first phase presentation formats. Each activity is represented by an icon and is positioned in chronological order on the line corresponding to its starting hour. This example illustrates the activities of user B on November 29, 2012.

subjects in finding trivial ways to distinguish days would be to show them only one day in the memorization phase. The subject would then have to recognize the same day among other days during the test phase. That way subjects would not be able to compare days for easy ways to differentiate them during the memorization phase. Moreover I could select for the test phase only days that are identical to the memorized day in trivial ways, such as getting up time and bedtime.

6.2.2 Task-based Evaluation with Domain Experts

Ultimately, the model will need to be tested in realistic scenarios in any projected application domain. To verify the capacity of a data-to-text system to benefit its users, a domain-specific task-based evaluation can be designed. Portet et al. (2009) is an example of this type of data-to-text system evaluation in the healthcare domain. In collaboration with domain experts, fictional scenarios presenting relevant anomalies could be elaborated. In domains such as healthcare where it would not be ethical to voluntarily reproduce anomalies, those scenarios would need to be simulated. Actors could be hired to enact those scenarios in a controlled environment or fictional data could be adapted from normal, existing real-life data. Throughout the simulation process domain expertise would be essential.

Once data containing known anomalies is obtained, the experiment can be planned. For comparison purposes, different presentation formats can be prepared: raw data, visual presentation, generated text, and human-written text. For human-written texts, again the collaboration of domain experts is needed. The data can then be presented to another set of domain experts representative of the projected users of the system. Finally, their performance at correctly identifying anomalies in the data for each presentation format can be measured. Results can be analyzed to see if the system is adequate and identify aspects that need to be improved.

Needless to say, domain expertise and recording of artificial scenarios are generally costly and this type of evaluation can be very expensive in both time and money.

6.3 Conclusion

This chapter first presented the method and results of an evaluation of the textual quality of the ADL reports. 13 judges evaluated together 32 generated narratives on 6 criteria. The evaluation revealed that the reports were considered highly understandable by the judges. Additionally, the lexical choice mechanics seem adapted to a report-style narrative. Flow between sentences, although not bad, was the generated texts' biggest weakness. Although it is not possible to compare those results directly with previous temporal data-to-text evaluations, the methodology is comparable and adequate considering the limited resources.

Next were sketched two extrinsic evaluation designs that could be used in the future to evaluate the validity of the model regarding the generated text's communicative goal. The first, which has already been subjected to a dry run, proposes memorization as a way of testing the communication of unusual events without actually having to define them. The second consists of an anomaly identification task that would require collaborating with domain experts.

The next chapter will conclude this thesis by presenting, among other things, ways in which the insights gained in the text quality evaluation could be exploited to improve the model.

CHAPTER 7

CONCLUSION

To conclude this thesis, Section 7.1 summarizes what has been presented: the assisted temporal data interpretation model, the extraction of the associative network, the generation of a narrative, and the evaluation of the model. Section 7.2 discusses future work stemming from lessons learned from the evaluation, from the proposed model's unexplored corners and combination with other approaches, and from considering more sources of data to experiment with.

7.1 Summary

Data containing information about events abounds in our technological society. 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. This thesis exploits this and proposes an assisted temporal data interpretation model by narrative generation.

First previous work serving as theoretical foundation is summarized. The notion of narrative and psychological principles such as the covariation principle, causal networks and causal chains are defined. Association rule mining is then formally characterized. After introducing a popular NLG pipeline architecture and the concept of rhetorical relations, several narrative NLG systems are presented. Some create fictional stories, others generate from real-life data. Narrative data-to-text systems seem to acknowledge causal relations as important. However, they play a secondary role in their document planners and their identification relies mostly on domain knowledge. To support evaluation design, some intrinsic and extrinsic NLG evaluations are also presented.

The thesis proceeds with the proposition of an assisted temporal data interpretation model by narrative generation. This model demonstrates that it is possible to structure a narrative with the help of 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 associations found do not correspond necessarily directly to causal relations. Rather they function as hints for the reader towards forming causal hypotheses. 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. This model should be applicable to any repetitive temporal data, preferably including actions or activities, such as Activity of Daily Living (ADL) data.

Before the data-to-text process can begin, association rules have to be gathered. Sequential association rules are obtained by computing the relative frequencies of certain patterns. Sequential association rule candidates are then selected based on the criteria of confidence and significance. Expected association rules have relatively high confidence and unexpected ones, very low confidence. World and domain knowledge rules are crafted manually by specifying how to measure the similarity of some aspect of a pair of events. They can also be based on causal patterns difficult to detect statistically in the training data.

The association rules can then be applied to the interpretation of a specific period to summarize. Pairs of events for which an association rule applies are associated. Those associations are marked with the corresponding association type and, in the case of sequential associations, the corresponding confidence. Some extra associations are then derived. Together the events and associations form an associative network.

Now that the associative network is available, natural language generation proper can begin. The most important step is document planning, comprising event selection and document structuring. For event selection, the model relies on the confidence of sequential associations. The goal is to select the most unusual facts. The assumption is that an event that is implied by another one with a relatively high probability may be excluded from the selected events and left implicit in the text.

Next comes the determination of the structure of the narrative, which is called the connecting associative thread. The role of this structure is to allow the reader to follow associations from the beginning to the end of the text. It takes the form of a spanning tree over the previously selected

associative sub-network. The associations it contains are selected based on association type preferences and relative temporal distance. The connecting associative thread is then segmented into paragraphs, sentences, and phrases and the associations are translated to rhetorical relations.

The NLG pipeline continues with microplanning. This step defines lexico-syntactic templates describing each event type. In the ADL implementation of the model, English and French versions of each lexico-syntactic template were hand-coded. When two event descriptions need to be assembled in the same sentence, a discourse marker expressing the specified rhetorical relation is employed. For the purpose of combining sentences, a main event and a preceding main event are determined for each sentence. When the associative thread parent of the main event is not the preceding main event, an anaphor is added to the sentence front discourse marker.

The last NLG step is surface realization. The ADL implementation used for the experiments uses the SimpleNLG-EnFr Java library as a surface realizer. This enables English/French bilingual generation when combined with the bilingual specifications provided by microplanning.

Because the generated text is about associated events, it can be interpreted as a narrative by the concerned human reader. This means that a causal representation is built in the mind of the reader with the help of their world and domain knowledge. The connecting associative thread can be followed throughout and give rise to causal hypotheses.

The proposed model has scientific value to the extent that it can be evaluated. NLG evaluations can be intrinsic or extrinsic. In an intrinsic evaluation, the textual quality of reports generated from the UCI ADL Dataset was rated by naive judges. The 32 narratives were evaluated on 6 criteria. The results show that the texts were quite understandable and the lexical choice mechanics adequate. Although not bad, flow between sentences could still be improved.

Two extrinsic designs are proposed to complete the model's evaluation. The first consists of a memorization experiment that tests that the generated narratives communicate unusual events while avoiding the problem of defining this notion. On the contrary, the second scheme relies on domain experts for the preparation of an anomaly identification task.

The next section suggests further research avenues, including ones informed by the analysis of the text quality evaluation results.

7.2 Future Work

An obvious starting point for further research would be to implement the two proposed extrinsic evaluation designs: the memorization experiment and the task-based evaluation with domain experts. Additionally, the results of the text quality evaluation could be leveraged. Flow between sentences was found to be a weak point of the reports in this evaluation. A possible way of improving it would be to modify document structuring such as to minimize discontinuities. According to the event-indexing model (Zwaan et al., 1995), sentence-reading times increase with the number of discontinuities in temporality, spatiality, protagonist, causality, or intentionality. Instead of selecting the best associations based on association preferences and relative temporal distance, the selection could consist in determining the transition with the least discontinuities. Exactly which of these and other discontinuity dimensions to take into account and how to quantify them are issues to be explored.

The mining of association rules is the first step in the model and could also be a starting point for future improvements. Chambers and Jurafsky (2008) learn narrative event chains (partially ordered sets of events with a common protagonist) from a news stories corpus. For this they use pointwise mutual information (PMI) to measure the relation between two events, instead of the probability of independence according to the binomial distribution. They then use a temporal classifier to determine a partial order. Finally, they cluster events using the PMI scores to form in effect undirected n-ary associations. Those could be converted to directed associations if confidence was also computed. An approach that combines the advantages of event chains and sequential association rules could be developed.

Section 4.1.2 introduced manually written association rules based on similarity and specific causal patterns. Only one type of similarity-based association rules was implemented as part of this thesis research and no specific causal pattern-based association rule. Further research will need to implement and test that part of the model.

In the Tag2Blog system (Ponnamperuma et al., 2013), probable contextual information is added with the help of domain knowledge. This may not be acceptable in every case, as some applications require that only confirmed information be included in the text. However, the model should still allow room for world and domain knowledge rules that add content or event details

and not just associations.

As evoked in Section 3.1.2, this model adopts a bottom-up approach to document planning that makes it in principle applicable to any domain. However, when the domain is known and the organization of the target text type can be rigorously analyzed, it may be better to specify the document structure in a more top-down fashion. That could mean that some topics would always be present in the same order or that the high-level rhetorical structure would be constrained. A possible avenue of investigation would be to determine how top-down constraints could be introduced in document planning while keeping a bottom-up approach for unconstrained parts. Note that having the text begin with the initial situation and end with the final situation can already be considered a top-down constraint.

How to determine the initial and final situations in every case is itself a problem not yet solved. For now, as discussed in Section 5.1.2.1.1, the model assumes that the type of period to summarize has been chosen so that the chronologically first and last events can play that role. This is possible when the data can be segmented beforehand into well-delimited cycles that seem natural to the reader. On the contrary, 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.

The question of the content of the initial and final situations for multivariate data has also been raised in Section 5.1.2.1.1. The beginning and end of the text should introduce and recall, respectively, the state of the variables relevant for its middle section, nothing more and nothing less (McKinlay et al., 2009). This would require modifying content selection in the model accordingly.

Section 5.2.1 explained how two distinct lexico-syntactic templates were needed to describe an event. This was because an event description could be a first mention or an anaphor. In the first case a clause was needed and in the second a noun phrase. The relation between the two syntactic structures could be better systematized if the input to surface realization was a semantic representation. A surface realizer taking such an input was employed by Wanner et al. (2010) and Bouayad-Agha et al. (2012). It was developed using the graph transducer workbench MATE (Bohnet et al., 2000). Lambrey and Lareau (2015) recently expanded its treatment of

collocations. With such a realizer, information on collocations (e.g. what support verb to use with each noun) could be moved to the lexicon. A semanticon could store the lexical entries corresponding to each meaning. The lexicon and semanticon could of course be personalized to fit the application domain and genre. This would constitute a more modular approach to paraphrasing and would help generate more varied texts.

Another avenue of research would be to vary the input temporal data. Texts could be generated from bigger ADL datasets, such as the CASAS datasets (Cook et al., 2013), or datasets belonging to other domains. Section 3.4 suggested some alternative sources of temporal data: mobile and wearable devices, medical sensors, surveillance videos, digital transactions, click-through data, and server logs. Error trace data such as the input to the system of Farrell et al. (2015) could be added to the list. It would be interesting to fine-tune all parameters for each of those to see if ideal values vary from domain to domain. Moreover, bigger datasets could enable limiting training data in experiments to what it should be, that is only the data preceding the period to summarize (this topic was discussed in Section 3.5). If more data were not enough to compensate for data sparsity, the idea of using some sort of smoothing method on sequential statistics could be explored.

Throughout this thesis the type of input data considered was temporal data. Section 4.1.2.1 introduced the idea of having associations based on similarity of location. Yet all physical events have not only a time of occurrence but also a location. Spatiotemporal data already constitute the input of some data-to-text systems (Baez Miranda et al., 2014, 2015, Ponnamperuma et al., 2013, Turner et al., 2008). Given spatiotemporal data, why not extend the use of data mining techniques to the extraction of not only sequential association rules but also distance-based association rules and even hybrid distance-based sequential association rules?

REFERENCES

Emily Ahn, Fabrizio Morbini, and Andrew Gordon. 2016. Improving fluency in narrative text generation with grammatical transformations and probabilistic parsing. In *Proceedings of the 9th International Natural Language Generation conference*, pages 70–73. Association for Computational Linguistics, Edinburgh, UK.

Daniel Athearn. 1994. *Scientific Nihilism: On the Loss and Recovery of Physical Explanation*. SUNY Press. Google-Books-ID: N2zXlDhqsAkC.

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

Belén A. Baez Miranda, Sybille Caffiau, Catherine Garbay, and François Portet. 2015. 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. Association for Computational Linguistics, Brighton, UK.

Mieke Bal. 2009. *Narratology : introduction to the theory of narrative*. University of Toronto Press, Toronto, third edition.

Anja Belz and Ehud Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In *EACL*.

Bernd Bohnet, Andreas Langjahr, and Leo Wanner. 2000. A Development Environment for an MTT-based Sentence Generator. In *Proceedings of the First International Conference on Natural Language Generation - Volume 14*, INLG '00, pages 260–263. Association for Computational Linguistics, Stroudsburg, PA, USA.

Nadjet Bouayad-Agha, Gerard Casamayor, Simon Mille, and Leo Wanner. 2012. Perspective-oriented Generation of Football Match Summaries: Old Tasks, New Challenges. *ACM Trans. Speech Lang. Process.*, 9(2):3:1–3:31.

Charles B. Callaway and James C. Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. In *ACL*, volume 94305, pages 789–797. Citeseer.

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

Keith Dowding. 2015. *The Philosophy and Methods of Political Science*. Palgrave Macmillan. Google-Books-ID: yyopCwAAQBAJ.

Rachel Farrell, Gordon Pace, and M Rosner. 2015. A Framework for the Generation of Computer System Diagnostics in Natural Language using Finite State Methods. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 52–56. Association for Computational Linguistics, Brighton, UK.

A. Fleury, M. Vacher, and N. Noury. 2010. SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):274–283.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93.

Gérard Genette. 2007. *Discours du récit*. Seuil, Paris.

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

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

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

Przemys\law Grzegorzewski. 2011. On the properties of probabilistic implications. In *Eurofuse 2011*, pages 67–78. Springer.

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

Catalina Hallett, Richard Power, and Donia Scott. 2006. Summarisation and visualisation of e-health data repositories. Nottingham, UK.

W. Hamalainen and M. Nykanen. 2008. Efficient Discovery of Statistically Significant Association Rules. In *ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 203–212.

James Hunter, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: Bt-nurse. *Artificial Intelligence in Medicine*, 56(3):157 – 172.

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

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

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

P. Lalanda, J. Bourcier, J. Bardin, and S. Chollet. 2010. Smart Home Systems. *Grenoble University, France*.

Florie Lambrey and François Lareau. 2015. Le traitement des collocations en génération de texte multilingue. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, pages 579–585. Association pour le Traitement Automatique des Langues, Caen, France.

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

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

Carlos León and Pablo Gervás. 2010. Towards a Black Box Approximation to Human Processing of Narratives Based on Heuristics over Surface Form. In *2010 AAAI Fall Symposium Series*.

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

Nicole Maslan, Melissa Roemmele, and Andrew S. Gordon. 2015. One Hundred Challenge Problems for Logical Formalizations of Commonsense Psychology. In *Proceedings of the Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense-2015)*. Stanford, CA.

A. McKinlay, C. McVittie, E. Reiter, Y. Freer, C. Sykes, and R. Logie. 2009. 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.

Marco Munstermann, Torsten Stevens, and Wolfram Luther. 2012. A Novel Human Autonomy Assessment System. *Sensors*, 12(6):7828–7854.

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

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

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

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

Kapila Ponnamperuma, Advaith Siddharthan, Cheng Zeng, Chris Mellish, and René van der Wal. 2013. 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. Association for Computational Linguistics, Sofia, Bulgaria.

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

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

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

Lauren J. Silbert, Christopher J. Honey, Erez Simony, David Poeppel, and Uri Hasson. 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, 111(43):E4687–E4696.

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

Douwe Terluin. 2008. *From Fabula to Fabulous: Using Discourse Structure Relations to Separate Paragraphs in Automatically Generated Stories*. Ph.D. thesis.

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

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

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

Ross Turner, Somayajulu Sripada, Ehud Reiter, and Ian P. Davy. 2008. Using Spatial Reference Frames to Generate Grounded Textual Summaries of Georeferenced Data. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 16–24. Association for Computational Linguistics, Stroudsburg, PA, USA.

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

Pierre-Luc Vaudry and Guy Lapalme. 2015. Narrative Generation from Extracted Associations. In *Proceedings of the 15th European Workshop on Natural Language Generation*, pages 136–145. Association for Computational Linguistics, Brighton, United Kingdom.

Pierre-Luc Vaudry and Guy Lapalme. 2016. Assembling Narratives with Associative Threads. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pages 1–10. Association for Computational Linguistics, Edinburgh, United Kingdom.

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

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

Appendix I

Text Quality Evaluation Form Sample

Text Quality Evaluation

Thank you for having volunteered to participate in this study. Your task is to evaluate the textual quality of five texts numbered A1, A9, B3, B11, and B19. The context and rating criteria are explained below and are repeated at the beginning of each section. Please be as objective as possible.

* Required

Section 1: Evaluation of text A1

Context: The texts you are evaluating are reports in the form of narratives that summarize the Activities of Daily Living (ADL) of a person in the course of one day. Each text reports on a different day for one of two users (user A or user B). The texts are based on real data. Examples of ADLs are sleeping, going to the toilet and eating. The reader of this report could be a health-care professional following this person to check for signs of dementia, for example. It is assumed that the reader is already familiar with the daily habits of the user.

Rating: Suppose you had to correct or rewrite the text in order to make it perfect. Please rate the report by specifying what proportion of each aspect of the text you could reuse as is. You will have to select a number from 0 to 5. Here is what each number means:

0. Too bad to be of any use. It would be better to start over from scratch.
1. Very few things can be kept. The rest must be rewritten.
2. A minority of this aspect of the text can be kept as is. Big changes must be made.
3. Most of this aspect of the text can be kept as is. Some changes need to be made.
4. Almost perfect. Some very minor changes.
5. Perfect. Nothing to improve.

Text A1

OrdonezA Monday, 28 November 2011 02:27 AM - Tuesday, 29 November 2011 11:31 AM

OrdonezA got up at 10:18 AM. Then he went to the toilet and took a shower. After his 10:25 AM shower, he ate his breakfast.

In addition to his 10:25 AM shower, he went to the toilet at 1:06 PM. Also as usual at 1:09 PM he left for 20 minutes.

Additionally at 2:22 PM he went to the toilet.

In addition as usual at 3:04 PM he groomed, spent 5 hours in the living room and then he had a snack.

Beside having gotten up at 10:18 AM, he went to bed as usual at 2:16 AM.

1. **Overall: What proportion of the text corresponds to how that kind of report should be in general?** *
   Mark only one oval.

	0	1	2	3	4	5	
Nothing is as it should be.	⬭	⬭	⬭	⬭	⬭	⬭	Everything is as it should be.

2. **Style: What proportion of the text is written in a style appropriate for that kind of report?** *
   Mark only one oval.

	0	1	2	3	4	5	
The style is bad throughout.	⬭	⬭	⬭	⬭	⬭	⬭	The style is appropriate throughout.

3. **Grammaticality: What proportion of sentences are grammatically correct?** *
   Mark only one oval.

	0	1	2	3	4	5	
All sentences are ungrammatical.	⬭	⬭	⬭	⬭	⬭	⬭	All sentences are grammatical.

4. **Flow: What proportion of sentences flow well from one to the next?** *
   Mark only one oval.

	0	1	2	3	4	5	
All sentence transitions are awkward.	⬭	⬭	⬭	⬭	⬭	⬭	All sentence transitions are smooth.

5. **Vocabulary: What proportion of word choices are appropriate?** *
   Mark only one oval.

	0	1	2	3	4	5	
All word choices are bad.	⬭	⬭	⬭	⬭	⬭	⬭	All word choices are appropriate.

6. **Understandability: What proportion of the text is perfectly understandable?** *
Mark only one oval.

	0	1	2	3	4	5	
The whole text is difficult to understand.	⬭	⬭	⬭	⬭	⬭	⬭	The whole text is perfectly understandable.

7. **Comments: Please feel free to leave any comments or explanations about your evaluation of this text**

.....

.....

.....

.....

.....

Section 2: Evaluation of text A9

Context: The texts you are evaluating are reports in the form of narratives that summarize the Activities of Daily Living (ADL) of a person in the course of one day. Each text reports on a different day for one of two users (user A or user B). The texts are based on real data. Examples of ADLs are sleeping, going to the toilet and eating. The reader of this report could be a health-care professional following this person to check for signs of dementia, for example. It is assumed that the reader is already familiar with the daily habits of the user.

Rating: Suppose you had to correct or rewrite the text in order to make it perfect. Please rate the report by specifying what proportion of each aspect of the text you could reuse as is. You will have to select a number from 0 to 5. Here is what each number means:

0. Too bad to be of any use. It would be better to start over from scratch.
1. Very few things can be kept. The rest must be rewritten.
2. A minority of this aspect of the text can be kept as is. Big changes must be made.
3. Most of this aspect of the text can be kept as is. Some changes need to be made.
4. Almost perfect. Some very minor changes.
5. Perfect. Nothing to improve.

Text A9

OrdonezA Tuesday, 6 December 2011 12:42 AM - Wednesday, 7 December 2011 10:45 AM

OrdonezA got up at 10:07 AM.

As usual at 11:12 AM he took a shower and then he ate his breakfast.

In addition at 5:13 PM he went to the toilet. Also at 6:07 PM he groomed, spent 1 hour in the living room and then he had a snack. Additionally at 7:25 PM he groomed.

Appendix II

Publications of the author related to this thesis

Below is a bibliography of the author's publications that are related to this thesis. The topics they cover are narrative data-to-text and bilingual surface realization. Publications cited in the thesis are marked with a star (*).

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

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

· Anne H. Schneider, Alasdair Mort, Chris Mellish, Ehud Reiter, Phil Wilson, and Pierre-Luc Vaudry. 2013. MIME-NLG Support for Complex and Unstable Pre-hospital Emergencies. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 198–199. Association for Computational Linguistics, Sofia, Bulgaria.

· Pierre-Luc Vaudry and Guy Lapalme. 2015. Causal networks as the backbone for temporal data-to-text. Paper presented at the First International Workshop on Data-to-Text Generation. Edinburgh, United Kingdom.

* Pierre-Luc Vaudry and Guy Lapalme. 2015. Narrative Generation from Extracted Associations. In *Proceedings of the 15th European Workshop on Natural Language Generation*, pages 136–145. Association for Computational Linguistics, Brighton, United Kingdom.

* Pierre-Luc Vaudry and Guy Lapalme. 2016. Assembling Narratives with Associative Threads. In *Proceedings of the INLG 2016 Workshop on Computational Creativity and Natural Language Generation*, pages 1–10. Association for Computational Linguistics, Edinburgh, United Kingdom.