Towards a Possibility-Theoretic Approach to Uncertainty in Medical Data Interpretation for Text Generation

François Portet[1] and Albert Gatt[2]

[1] Laboratoire d'Informatique de Grenoble, Grenoble Institute of Technology, France
francois.portet@imag.fr
[2] Institute of Linguistics, University of Malta, Malta
albert.gatt@um.edu.mt

Abstract. Many real-world applications that reason about events obtained from raw data must deal with the problem of temporal uncertainty, which arises due to error or inaccuracy in data. Uncertainty also compromises reasoning where relationships between events need to be inferred. This paper discusses an approach to dealing with uncertainty in temporal and causal relations using Possibility Theory, focusing on a family of medical decision support systems that aim to generate textual summaries from raw patient data in a Neonatal Intensive Care Unit. We describe a framework to capture temporal uncertainty and to express it in generated texts by mean of linguistic modifiers. These modifiers have been chosen based on a human experiment testing the association between subjective certainty about a proposition and the participants' way of verbalising it.

1 Introduction

Clinical decision support systems (CDSS) run into problems when there is temporal uncertainty or inaccuracy in their input data, which can arise for a variety of reasons. For example, medical staff often record events when they have time to do so, rather than when they actually happened. In addition, existing database management systems tend not to deal with temporal data in a principled fashion [1]. Uncertainty and inaccuracy make the tasks of reasoning about temporal and causal relationships more difficult, especially where input data is provided in a raw form.

Classical CDSS (especially expert systems) typically approach the problem of uncertainty either by restricting output to what the system is completely certain about, or by communicating findings using a ranking mechanism [2]. In contrast to such systems, the family of CDSS being developed in the BabyTalk project [3] aim to provide *textual* summaries of heterogeneous medical data (both automatically and manually entered) to support decisions by carers in Neonatal Intensive Care Units (NICUs). The goal is to communicate the relevant aspects of patient data, by using Natural Language Generation (NLG) techniques to produce a descriptive summary, leaving it up to the user to decide on the best course of action. The emphasis on generating textual summaries contrasts with current approaches to medical decision support, which mainly rely on visualisation techniques to present data. A recent off-ward evaluation of a prototype system, BT-45, suggested that textual summarisation is at least as effective in supporting decisions as

D. Riaño et al. (Eds.): KR4HC 2009, LNAI 5943, pp. 155–168, 2010.

current visualisation techniques [3]. However, the robustness and effectiveness of such systems depend on the extent to which they incorporate a principled approach to temporal representation and uncertainty. To be an effective decision support tool, a summary must permit the reader to reconstruct the temporal sequence of the events that it narrates, and make clear the relations between them. Furthermore, where the precise time at which events occurred is not available, the inference of temporal and/or causal relationships between events can be compromised. The resulting uncertainty should arguably be reflected in the texts produced. Failure to do this can result in erroneous decision-making, whose consequences can be serious given the fault-critical nature of the environment in the NICU.

In this paper, we describe an approach to dealing with temporal uncertainty in the reasoning component of these systems based on Possibility Theory. We also discuss how the outcome of reasoning about temporal relations, and other relations that are contingent upon them, such as causality, can be exploited in an NLG component to communicate uncertainty in the data using expressions such as epistemic modals (e.g. *may* and *must*). The approach is intended to be generalisable to some extent to those kinds of situations in which raw input data needs to be processed prior to carrying out any form of reasoning, and the data itself contains incorrect or uncertain times for the events under consideration. On the other hand, as emphasised above, medical decision support is a particularly important domain in which to deal with these issues, both because of the high density of the data being processed (resulting in increased likelihood of temporal inaccuracy) and the potential consequences of failing to deal with uncertainty.

In the rest of this paper, we first begin by reviewing some related work (Section 2), followed by a motivating example from the NICU domain (Section 3). We then introduce our formalism, with a focus on uncertain relations between intervals (Section 4). Section 5 describes the approach to reasoning with uncertain temporal information. Section 6 discusses how uncertainty in temporal and causal relations can be used to inform the choices made by a text generator in producing a summary, with particular emphasis on the use of modal expressions. Finally, Section 7 reports preliminary results from a web-based experiment testing the association between subjective certainty about a proposition and the participants' choice of linguistic expressions to convey it. We conclude with remarks on future work in Section 8.

2 Related Work

Temporal reasoning is crucial to temporal abstraction [4]. In medicine, it is especially important in making inferences for diagnosis, recommendations based on computerized guidelines, or textual summarisation. Formalisms for temporal reasoning typically rely on the use of temporal constraints [5,6,7,8,9,10] which can either be qualitative (for example, Allen's temporal relations [5]) or numerical [11] (such as a range of temporal distances between two events). Both representations account for different kinds of imprecision: the former is suitable for relations such as *A is after B*; the latter for expressing relations such as *the temporal distance between A and B is between 2 and 4 hours*. With imprecise data and knowledge (which often occurs if the source is a human agent), reasoning leads to uncertainty in temporal relations, which many formalisms represent either through the use of probability distributions [7] or fuzzy sets [6,8].

Regarding probability based reasoning, Ryabov and Trudel [7] have proposed qualitative probabilistic temporal interval networks in which relations between intervals are labelled with a probability value. While the framework supports reasoning with uncertain relations, it presupposes that the probabilities are known. This strong assumption limits the model's applicability in domains where many concepts are to be dealt with (such as the NICU) and precise estimates of probabilities cannot be made due to the absence of large volumes of annotated data. Moreover, probability theory lacks the flexibility to express partial ignorance. Indeed, when the probability of a proposition A is known, the probability of its complement is fully determined ($P(\bar{A}) = 1 - P(A)$). In contrast, non-classical formal theories, such as Possibility Theory or the Dempster-Shafer Theory (DST), deal with partial ignorance by representing uncertainty using two complementary measures (sometimes referred to as upper and lower probabilities).

A recent example of a qualitative temporal reasoning formalism based on fuzzy sets is Badaloni *et al.*'s IA^{fuz} framework [8], which expresses uncertainty via constraints whose priorities correspond to degrees of plausibility. However, this approach does not address uncertainty derived from data (e.g. inaccurately timestamped events). To deal with vagueness in data, Vila and Godo [12] generalised the classical Temporal Constraint Satisfaction Problem (TCSP) by considering temporal constraints defined by fuzzy sets. This approach has been used in the medical domain for ICU diagnosis [9], and for the processing of clinical texts [10].

More recently, Dubois *et al.* [6] have proposed a framework for reasoning with a fuzzy version of Allen's temporal relations. This formalism is of linguistic relevance, insofar as there is an intuitive mapping between the representation of graded vagueness in temporal relations and vague linguistic operators (e.g. *'approximately' equal*), as well as expressions about the certainty that these relations hold (e.g. *A may have happened shortly before B*). Although our focus in this paper is on uncertainty, we are also interested in extending our approach to deal with vagueness (which arises, for example, in the classification of temporal relations as 'shortly after'). Another attractive property of the formalism is that it derives the uncertainty of temporal relations directly from the intervals over which events occur, by representing these intervals as possibility distributions. This makes the approach amenable to a direct application to raw data. Furthermore, Possibility Theory has been shown to better reflect the qualitative nature of human reasoning with uncertainty [13] and is thus better adapted to the communication of uncertainty in natural language for decision making than other formalisms (including DST and other two-level approaches that are exclusively numerical). For these reasons, we build our approach on the basis of this work.

3 A Motivating Example

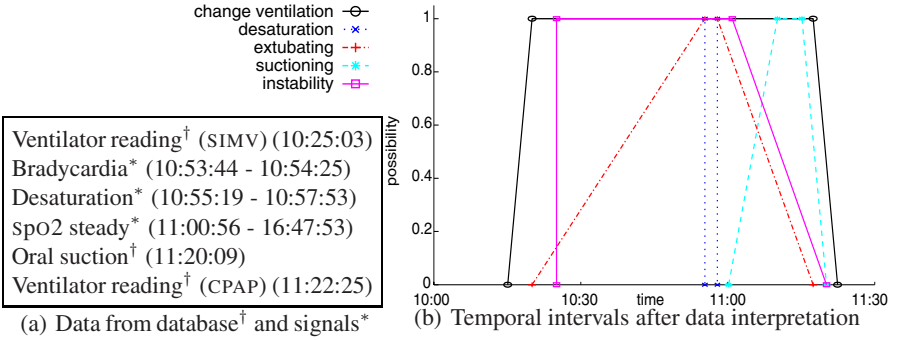
NICU data is of two kinds: 1) discrete records logged by the medical staff on the NICU database, such as drug administration; 2) physiological data sampled at high frequency from probes measuring heart rate (HR), oxygen saturation (SaO2 or SpO2), etc. We assume an interval-based representation. The systems under discussion [3] process the input data in four stages: A *data analysis* stage identifies significant trends and patterns in the physiological data, as well as data records in the database, mapping them to

concepts in a domain-specific ontology. Then, *data interpretation* infers temporal and causal relations between events. Subsequently, the NLG stage selects important events, plans the structure of the summary (*document planning*) and maps the selected events and their relations first to semantic representations (*microplanning*) and finally to syntax (*realisation*). Our focus in this paper is on extensions to data interpretation and microplanning, to deal with uncertainty in temporal relations in reasoning and natural language semantics.

To make the problem concrete, consider the sample of data in Figure 1(a), consisting both of events logged in the database by NICU staff and patterns automatically discovered during signal analysis. A corresponding fragment of a nurse shift summary, written by an experienced neonatal nurse, is shown below.

Example 1. He is currently on nasal CPAP in air, having been extubated today [...] Prior to extubation his SpO2 and HR showed compromise during handling with desaturations and HR decelerations

The example highlights a number of possible sources of uncertainty for a system that (unlike a nurse) is entirely reliant on recorded data. The extubation has not been recorded in the database and must be inferred. There are two relevant facts. First, the baby's having been on SIMV ventilation (a type of ventilation requiring the patient to be intubated, i.e. to have a tube in her throat supplying air) earlier is consistent with her having been intubated. Second, the change to CPAP (a kind of ventilation less severe than SIMV) indicates that the baby has been extubated. The precise time at which the extubation was carried out is, however, uncertain, and there is no specific event corresponding to the placement of the baby on CPAP. The latter too is inferred from the two consecutive ventilator readings, the second of which shows a change in ventilation mode. Thus, the precise time of the ventilator change event itself cannot be determined, though it must have overlapped with the extubation. Finally, the text makes reference to instability in heart rate (HR) and oxygen saturation (SpO2). This is an abstraction that the system needs to perform from the signals. Once again, the interval over which the period of instability holds is fuzzy. Moreover, the reference to 'handling' suggests that the instability occurred during the extubation and prior to its completion.



(a) Data from database† and signals*

(b) Temporal intervals after data interpretation

Fig. 1. Data from database and signal analysis and its temporal representation after interpretation

The outcome of data interpretation is shown in Figure 1(b). The trapezoidal representation of event intervals indicates the inaccuracy in the time at which they started and ended. This affects the certainty with which relationships between them can be inferred, such as the temporal overlap between the ventilator change and the extubation, and the causal relation inferred between the extubation and the instability period, which is supported by the knowledge that handling must have occurred during the extubation, and that it often causes temporary disruptions in a patient's physiological parameters. Though domain knowledge reduces uncertainty, it is not always possible to eliminate it. In such cases, a summary should communicate this uncertainty to the human reader.

4 Representing Temporal Information

The formalism used to represent intervals and uncertain relations between them is built on Possibility Theory as formalised by Dubois *et al.* [6]. In what follows, lowercase italic letters (*a,b,c,...*) denote dates and normal uppercase (A,B,C,...) inaccurate intervals. Recall that in Possibility Theory, the uncertainty about an inaccurate interval A that holds at date $d \in \mathbf{Z}$ can be evaluated by the dual measures of possibility II and necessity (also called *certainty*) N, as follows:

$$II(\mathrm{A}(d)) = h_{\mathrm{A}}(d) \tag{1}$$

$$N(\mathrm{A}(d)) = 1 - h_{\bar{\mathrm{A}}}(d) \tag{2}$$

Where $h_{\mathrm{A}} \in [0, 1]$ is the hold function of the interval A, representing the degree to which A has possibly occurred at date d, and $\bar{\mathrm{A}}$ is the complement of A. Additionally, $II(\mathrm{A}) = \max h_{\mathrm{A}}$, $II(\mathrm{A}(d) \vee \mathrm{B}(d)) = \max\{II(\mathrm{A}(d)), II(\mathrm{B}(d))\}$, $N(\mathrm{A}(d) \wedge \mathrm{B}(d)) = \min\{N(\mathrm{A}(d)), N(\mathrm{B}(d))\}$. Moreover, by formula 2 and the following property:

$$II(\mathrm{A} \cup \bar{\mathrm{A}}) = 1 \tag{3}$$

the necessity of an interval A at date d can be summarised as $\mathrm{A}(d)$ *is certain only if no interval contradicting* A *(i.e.,* $\bar{\mathrm{A}}$*) is possible at time* d. If several contradictory intervals are completely possible at the same time (e.g., several mutually exclusive values of a device setting), no certainty exists. In the following, we restrict ourselves to trapezoidal hold functions. Thus, an inaccurate interval is given by the following definition:

Definition 1 (inaccurate interval). *An inaccurate interval* A *is a 5-tuple* $\langle o, s, e, \alpha, \beta \rangle$*, where* $o \in \mathcal{O}$*;* $\mathcal{O}$ *is the domain of concepts;* $s, e, \alpha, \beta \in \mathbf{Z}$*,* $s - \alpha \leq s$*,* $s \leq e$*, and* $e \leq e + \beta$*.*

A is seen as a concept with a trapezoidal fuzzy set that describes the period during which A possibly holds. The $[s, e]$ interval is the core of the fuzzy set (i.e. the latest possible start and the earliest possible end) and $[s - \alpha, e + \beta]$ is the support (i.e. the earliest possible start and the latest possible end). In what follows, we reserve the term 'interval' for inaccurate intervals. A trapezoid function leads to $II(\mathrm{A}) = 1$, since $\max h_{\mathrm{A}} = 1$. Thus, it is completely possible that A holds, though its start and end time are not known with certainty.

In our application domain, the 'meaning' that an interval conveys can usually be determined by linking each record to a concept in the knowledge representation, which in the present case takes the form of an ontology. One of the uses of associating intervals with concepts is that domain knowledge can then constrain or validate some intervals. A full discussion would take us beyond the scope of this paper, but to summarise, temporal knowledge such as max or min duration or qualitative or quantitative ordering (e.g. A is usually followed by B within a few hours) can be used to remove ambiguity. This kind of expert knowledge is easier to embed into an ontology using fuzzy sets and logic than using probabilistic models. In the following, we distinguish the notion of *event* from the notion of *state*. Typically the former is related to actions or occurrences (e.g. therapy change, bradycardia) over short periods, while the later is related to a conditions which tend to persist over time unless some event perturbs them. Both notions can be represented by an inaccurate interval.

4.1 Temporal Relations between Intervals

Approaches to temporal reasoning [6,7] are usually based on some subset of Allen's 13 relations [5]. Here, we consider only the three disjoint temporal relations *before, intersects, after* (where *intersects* is related to, though different from, Allen's *overlap*) with the aim of dealing with the (un)certainty of temporal relations between intervals (e.g. what is the certainty that A is before B?). Consider the intervals A and B. The necessity that the end of A is before the start of B is given by [6]:

$$N_{es}(\text{A,B}) = 1 - \max_{b \leq a \in \mathbf{Z}} \{L(\text{B}.s, \text{A}.e), \min\{h_{\text{A}}(a), h_{\text{B}}(b)\}\} \tag{4}$$

where $L(x, y) = 1$ if $x \leq y$ and $L(x, y) = 0$ if $x > y$. Thus, the necessity of the end of A occurring before the start of B is the dual of the possibility that the start of B is before the end of A. L is used to constrain $N_{es}(\text{A,B})$ to be 1 when the core of A overlaps the core of B (in which case the possibility that the beginning of B is before the end of A is 1). Similarly, we can define the necessity $N_{ee}(\text{A,B})$ that the end of A is before the end of B, the necessity $N_{ss}(\text{A,B})$ that the start of A is before the start of B, and the necessity $N_{se}(\text{A,B})$ that the start of A is before the end of B. For intervals A and B, we define the three basic relations as follows:

$$N(\text{A before B}) = N_{es}(\text{A,B}) \tag{5}$$

$$N(\text{A after B}) = N_{es}(\text{B,A}) \tag{6}$$

$$N(\text{A intersects B}) = \min\{N_{se}(\text{B,A}), N_{se}(\text{A,B})\} \tag{7}$$

Note that the system described here does not use these relations to maintain a temporal network. Rather, temporal relations are established exclusively on the basis of observation, obviating the need for temporal constraint propagation. This is advantageous, given that the number of intervals can run into thousands, which would compromise the runtime efficiency of such a system were it to attempt to maintain a fully consistent

network. Solutions to the latter problem are known to have an exponential worst-case complexity [5]. However, it is precisely because of the reliance on direct observation that errors or uncertainties in the input data need to be rectified at the reasoning stage.

5 Temporal Abstraction and Interpretation

Interpretation is done in two stages: 1) application of *a priori* domain knowledge to independent intervals to alter their fuzzy sets to reduce ambiguity; and 2) application of temporal reasoning to abstract and interpret states and events. Our knowledge base consists of a large ontology developed within the BabyTalk project [3] (which contains more than 900 concepts) and rules acquired from interviews with experts.

For the first phase, when the intervals are first read in from the data source, knowledge can be deployed to directly constrain uncertainty about their temporal information. For example, the ventilator mode readings (i.e., SIMV and CPAP) in Figure 1 are not recorded with accurate timestamps (they are logged on an hourly basis). Due to property (3), the possibility of each of the ventilation mode values should be defined between these two readings. Theoretically, it could take any value in its domain. Since any value for the ventilator reading is possible at any time in principle, no certainty of any kind can be derived from these data alone, without reasoning based on the application of reasonable *a priori* constraints. These constraints consist in:

1. assuming persistence for states, unless there is evidence that a state has ceased to hold;
2. assuming that any state A is expanded by $delay$, to take into account a minimal delay between the human observation and the human recording on the computer;
3. assuming that any event A has actually happened before the transaction date (that is, the date at which the record for A is entered).

These simple rules are crucial for disambiguation. The first constraint enables the aggregation of intervals representing states with the same properties. The second one accounts for inertia in the recording of data (i.e. a value is only recorded if it has held true for a certain delay period around the transaction date). The third one is known to be typically true based on consultation with domain experts. Apart from these rules, knowledge encoded in the ontology (such as max and min duration) and in the expert rules, permits the modification of the fuzzy set of the interval. For the example in Figure 1(b), the baby is known to be intubated; however, the CPAP reading contradicts this, since the domain knowledge specifies that this kind of ventilation support *requires* the baby to not be intubated. Thus, the system infers that the ventilation mode has been changed over the period C and that an extubation event E has possibly existed between the two readings. The exact location of this extubation is still vague but again the knowledge base informs us that extubation can cause perturbation on the physiological signals (due, among other things, to handling). Thus, the maximal possibility for this extubation occurs during the period of the desaturation D, which is intersected by E. This reasoning explains the shape of E in Figure 1(b), which is completely possible during D, less possible during the change of ventilation period, and impossible otherwise. These outcomes are clearly strongly dependent on domain assumptions, but this

cannot be neglected, as basic knowledge of the domain is often an easy and reliable way to reduce complexity in reasoning.

After the first stage, the following values can be computed: $N(\text{C after E}) = 0$, and $N(\text{C intersects E}) = 1$. For E and the oral suction O, $N(\text{E before O}) = 0.42$, $N(\text{E after O}) = 0$, $N(\text{E intersects O}) = 0.58$. In addition, the following relationships are computed between the oral suction event O and the period of instability U in HR and SpO2: $N(\text{O intersects U}) = 0.68$, $N(\text{O after U}) = 0.32$.

Thus, although the temporal order of oral suction and extubation could not be established, it is more certain that suction has been performed right after or during extubation than before. This also explains part of the motivation behind the suction event.

Finally, inference rules are applied for abstraction and interpretation. In this framework, the validity of an inference chain is measured by its weakest links, so that the weight of a conclusion should be the weakest among the weights of its premises. As noted earlier, our reasoning is data-driven in the sense that the temporal relations considered are all and only those derived from data. As an example, the following rule is fired when an intervention (such as an extubation E or an oral suction O) intersects with an instability period (U), which represents the degree of variation of the physiological parameters related to respiration over periods of time. These periods are delimited by the main respiratory interventions.

Example 2. E is-a RESPIRATORY INTERVENTION $\wedge$ U is-a INSTABILITY $\wedge N(\text{E intersects U}) \geq \psi \Rightarrow N(\text{E causes U}) = N(\text{E intersects U})$

This rule matches the extubation and the oral suction in the example and infers that *O causes U* with necessity 0.68, while *E causes U* with necessity 1.

6 From Events to Text

In this section, we shall be concerned with the implications of the foregoing discussion for communicating uncertainty in the microplanning component. This takes as input a document plan — a labelled graph whose nodes are intervals or sequences of events, and whose edges are relations between intervals — and produces a semantic representation for the intervals in the document plan, which is then mapped to a syntactic representation. A partial document plan for our example is displayed in Figure 2, where the edge labels indicate the necessity with which a relation holds.

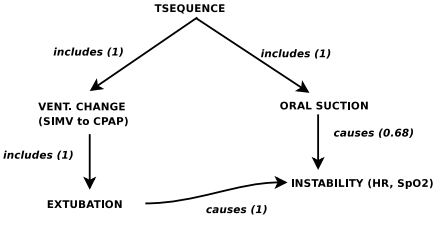


Fig. 2. Document plan fragment. Numbers in parentheses are necessity values.

One of the mechanisms that natural language provides for the expression of different degrees of (un)certainty is modality. Classical treatments of the semantics of modal expressions such as *can*, *must* and *may*, rely on their *modal force* (the degree of necessity or possibility expressed by the modal expression) and the *contextual background* against which they are interpreted. In the present case, our focus shall be on epistemic modality, where the relevant background is the speaker's knowledge. A proposition such as *x must/may have occurred* is roughly paraphrasable as *x must/may have occurred* in view of what is known [14]. Assuming, following Grice [15], that a speaker will not impart information beyond what is required unless it is relevant, qualifying an assertion in this way (e.g. *The extubation* may *have caused instability*) signals to the hearer/reader that the degree of a speaker's certainty is relevant to how the truth of the proposition should be evaluated [16]. This is particularly relevant for the present domain, where expressing uncertainty explicitly may alter the course of decision-making by a reader. It seems likely that this is also true of *must*. Although this has traditionally been taken as expressing logical necessity [14], the use of *must* suggests that the relativisation to the speaker's knowledge is important; this seems to be part of the pragmatic import of the use of the modal, and is compatible with the Gricean argument outlined above.

Typically, formal semantic treatments of modals are couched in a possible worlds framework [14,16]; this has also been adopted in NLG by Klabunde [17] to deal with (deontic) modality in a system that generates recommendations for course choices to students. In contrast to this work, the present approach proposes to view epistemic modal expressions as involving a direct mapping from different degrees of necessity or certainty (which reflects the epistemic 'modal force' of the proposition to be communicated) to linguistic expressions. One potential advantage of this approach is that, just as the necessity and possibility computations discussed in Section 4 are derived directly from data, so the use of epistemic modals is grounded in the data that constitutes the speaker's (system's) knowledge state.

To deal with modality in this way, we make the following assumptions about the lexical resources available to the microplanner. First, every relation between intervals in a document plan maps to a linguistic expression in the lexicon. For example, a *cause* relation maps to the verb *cause*. Modal auxiliaries are represented in the lexicon via a function $\mu : N(R) \rightarrow \text{AUX}$, which maps the necessity value of a relation to an epistemic modal auxiliary verb. A possible implementation of this function is sketched out below:

$$\mu(N(R)) = \begin{cases} \textit{may} \text{ if } l < N(R) \leq 0.6 \\ \textit{must} \text{ if } 0.6 < N(R) < 1 \\ \perp \text{ otherwise.} \end{cases} \tag{8}$$

where l is a lower bound on the certainty below which the relation is not expressed at all because it is too uncertain, and $\perp$ is `null`. By this formulation, a relation such as *cause* will be expressed with no qualification ($\perp$) if the certainty is 1, but may be qualified using *may* (which carries weak epistemic modal force) or *must* otherwise. Another possibility for expressing high degrees of certainty is *should*. However, a sentence such as *X should have caused Y* may be interpreted as implying violated expectation (i.e., *X was expected to have caused Y but didn't*). If this is the case, then using *should* would take the system's generated text beyond description and into something akin to

recommendation, since pointing out violated expectation may lead to an increased focus on the reader's part to check whether something went wrong.

The document plan in Figure 2 contains an additional complication: there are two events possibly contributing to the instability event, with different degrees of certainty. Here, there are two possibilities. If the extubation and suction events are aggregated, to form a single clause as in Example 3, then the certainty of their joint causal role in producing the instability is once again the weakest link in the causal chain (the minimum certainty value), following the reasoning adopted in Section 5.

Example 3. The baby was moved from SIMV to CPAP. He was extubated and underwent oral suction. This must have caused the instability in HR and SpO2.

An alternative strategy is to realise each clause separately, making the causal link explicit in each case. In the case of extubation, where the certainty is 1, no modal is used. In the second case, the assertion of causality is qualified via *must*.

Example 4. The baby was moved from SIMV to CPAP. He was extubated, causing the instability in HR and SpO2. He underwent suction. This too must have caused instability.

The best choice between these two alternatives is an open empirical question.

7 Empirically Grounding the Linguistic Model

The above illustration of how necessity values can inform lexical choice in microplanning is couched in largely intuitive terms. However, it throws up a number of questions which we are currently investigating. Among the relevant issues is the degree of certainty with which people interpret different modal expressions in epistemic contexts, as well as the other inferences that they generate. An answer to this question would serve as the basis for empirically grounding the lexical resources used by the system, as well as testing our intuitions regarding the different modal force of different expressions.

In order to answer these questions, we ran an experiment aimed at investigating the degree of certainty with which propositions describing simple events are interpreted by human speakers depending on the degree of temporal uncertainty associated with the events. Another aim of the experiment was to investigate speakers' choice of linguistic expressions to express uncertainty, with a view to incorporating this into our model of lexical choice. The linguistic expressions considered fall into three classes: 1) Epistemic modals (*must* and *may*), which are the focus of the previous section; 2) Adverbs of possibility (*possibly* and *perhaps*), which offer an alternative way of expressing uncertainty and were included for comparison; and 3) Negation (that is, sentences of the form *it is not the case that E occurred at t*).

7.1 Materials, Design and Procedure

The experiment was conducted over the web. Participants were shown a series of scenarios, each of which consisted of a background text and two temporally grounded propositions (**S1** and **S2**) describing two events (**E1** and **E2**), whose timing could be precisely or inaccurately known (see the top of Figure 3). The scenarios were designed

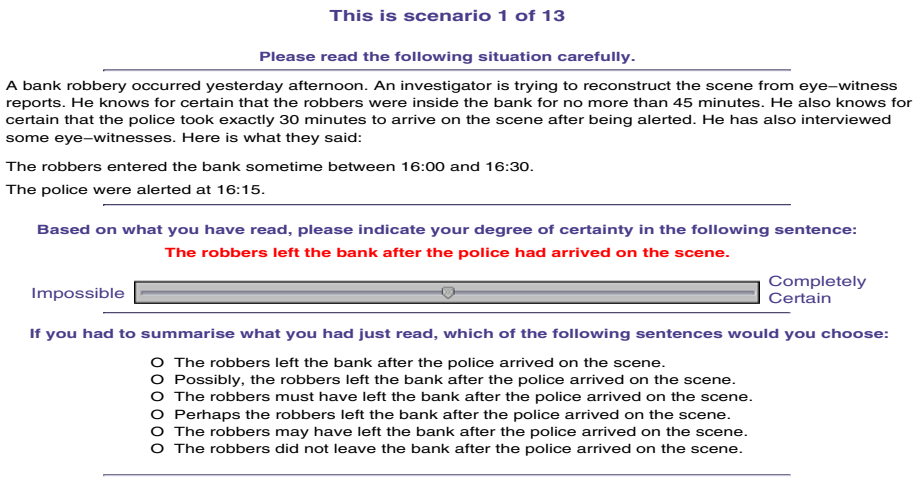


Fig. 3. Screenshot of one of the thirteen scenarios shown to the participants sequentially

to make it explicit that the events themselves actually happened for certain and that uncertainty was only related to their timing.

The experiment manipulated two factors. *Uncertainty* (3 levels) manipulated the extent to which the two events were precisely located in time. In the *no uncertainty* case, event times were expressed with a crisp value (e.g., *The robbers entered the bank* at 16:00.); in the *1-uncertainty* case, **E1** was expressed with a fuzzy temporal interval (e.g., *The robbers entered the bank* sometime between 16:00 and 16:30.); in the *2-uncertainty* case, both events had fuzzy temporal intervals. This factor enabled us to control the degree of certainty of temporal relations between events. The second factor, *Proposition Type* (4 levels), manipulated the type of proposition whose subjective certainty participants were asked to judge, namely: a simple proposition describing either **E1** or **E2**; or a compound proposition describing the temporal relation between the two events using one of the temporal connectives *before, after*, or *during*. Once participants had read a background text, they were asked to perform two tasks:

– **Judgement:** Given a certain scenario involving two events, participants were asked to judge their certainty that an event happened at a certain time or in a certain temporal order in relation to another event. Certainty was judged using a slider (see Figure 3 middle) representing the 'Ψ-scale' [13], which combines both possibility and necessity measures and ranges from 'impossible' to 'completely certain'. From the Ψ measure, the corresponding possibility II and necessity N can easily be reconstructed using (9).

$$II(P) = \begin{cases} 2 * \Psi & \text{if } \Psi \leq 0.5 \\ 1 & \text{if } \Psi > 0.5 \end{cases}, \; N(P) = \begin{cases} 0 & \text{if } \Psi \leq 0.5 \\ 2 * \Psi - 1 & \text{if } \Psi > 0.5 \end{cases} \tag{9}$$

– **Forced choice:** After judging the certainty of the proposition, they were asked to select, from among a set of sentences, the one they thought was most appropriate

to describe the temporal features of the scenario. These sentences represented the same proposition whose certainty they had judged, with or without expressions mitigating the temporal certainty (i.e. the simple proposition, corresponding to $\perp$ in (8) above, or propositions using *may*, *must*, *possibly*, *perhaps*, and a negated version of the proposition). The order in which propositions were presented was randomised for each scenario and participant.

Thirteen scenarios such as the one in Figure 3 were constructed. For each one, a version corresponding to each of the 13 combinations of *Uncertainty* and *Proposition Type* was developed (2 simple propositions, *S1* and *S2*, with and without uncertainty + $\{$*before,after,during*$\} \times \{$*no uncertainty, 1-uncertain, 2-uncertain*$\}$). A Latin Square Design was used to create 13 sets of items such that, within each set, each scenario occurred once in each condition, and no scenario occurred more than once in that condition across sets. Thirteen native speakers of English, all of them members of staff or postgraduate students at the University of Aberdeen, participated voluntarily in the experiment.

7.2 Results

Separate univariate ANOVAs were conducted to test the effect of *Uncertainty* and *Proposition Type* on the Ψ score, as well as on the possibility and necessity values. There was a significant main effect of *Uncertainty* on both Ψ ($F(2, 158) = 8.657; p < .001$) and on the derived necessity values ($F(2, 158) = 20.145; p < .001$), but not on possibility ($F(2, 158) = 0.003; ns$). Proposition Type exerted no main effect and there was no interaction between the two factors. This suggests that judgements of subjective certainty are strongly influenced by the manipulation of temporal uncertainty in the propositions being judged; however, it is *necessity* rather than possibility, which is the key correlate of these judgements. Thus, our focus on necessity values in the linguistic model sketched out in Section 6 has some prima facie justification.

Table 1 displays the mean Ψ, possibility and necessity values associated with the different types of propositions, based on participants' choices. In the default ($\perp$) case, both necessity and possibility are high. This is expected, given that, in natural language, unqualified assertions tend to be made in case subjective certainty is high. For *may* and *must*, the associated possibility values are high in both cases. On the other hand, there is a clear difference between sentences qualified with such modals and the default case: the latter is associated with a higher necessity value. Moreover, *may* involves lower necessity than *must*, as expected. As for adverbials and negation, necessity tends to be much lower, but the differences in possibility are clearer than in the case of modals.

A stepwise multinomial regression analysis to test the significance of II and N in determining the category of phrase selected by participants revealed that both II and N play a significant role in determining phrase choice (II: $\chi^2 = 61.8; p < .001$; N: $\chi^2 = 53.89; p < .001$) . Given the differences in the mean II and N values for modals on the one hand, and adverbials and negation on the other, separate regression analyses were conducted to identify the role of the two measures on the choice within either class. Interestingly, the models suggest a dissociation between necessity and possibility. Within the class of the two modals (together with the default $\perp$ case), regression showed

Table 1. Mean and standard deviations for subjective certainty (Ψ) and corresponding necessity and possibility values, as a function of phrase choice

	$\perp$	*must*	*may*	*perhaps*	*possibly*	*negation*
Ψ	.94 (.1)	.86 (.17)	.49 (.15)	.59 (.13)	.48 (.23)	.13 (.25)
possibility (II)	1 (0)	1 (.01)	.89 (.19)	.97 (.06)	.83 (.33)	.21 (.35)
necessity (N)	.88 (.2)	.71 (.34)	.09 (.19)	.21 (.23)	.13 (.24)	.05 (.20)

a significant effect of necessity ($\chi^2 = 36.07; p < .001$), but not possibility ($\chi^2 = 4.3; p > .1$). In contrast, the choice between the two adverbials and negation showed a significant role of II ($\chi^2 = 28.31; p < .001$) but not N ($\chi^2 = 1.12; p > .5$).

To summarise, the judgement results indicate that necessity is the primary correlate of people's subjective certainty judgements, but the phrase choice data suggests a dual role for possibility and necessity values. Specifically, there seems to be a dissociation between epistemic modals on the one hand, and adverbials of possibility (and negation) on the other. One possibility is that these linguistic expressions express different factors contributing to overall subjective certainty.

Although the results suggest that the intuitions underlying the linguistic model presented in Section 6 are on the right track, the mapping from necessity values to modals can be fine-tuned on the basis of this data. Subjects seemed to be tolerant of *some* degree of subjective uncertainty in opting for a non-qualified utterance ($\perp$), whereas our original proposal was to use $\perp$ only when $N = 1$. On the other hand, *must* is selected in cases where subjective certainty is quite high, and the gap between *must* and *may* is substantial. This still leaves open the question of the role of possibility, particularly of its apparent predictive power for the choice of adverbials and/or negation.

Though they are encouraging, the above results are preliminary, both in the sense that they are based on a relatively small pool of 13 subjects, and because they afford more sophisticated analysis. In addition to gathering more data, our ongoing work is addressing the question whether the subjective (un)certainty of two propositions linked by a temporal relation can be predicted from that of the simple propositions.

8 Conclusions and Future Work

This paper has proposed a possibility theoretic approach to the representation of inaccurate intervals and the knowledge-based discovery of temporal relations. The expression of these in text generation uses the uncertainty measure for qualifying relations via epistemic modal expressions. Our formalism is attractive in that it can be used to combine information from different sources (e.g., pattern recognition outputs, database entries, information extracted from free-text) while linguistic expressions are also directly grounded in measures of certainty based on the available information. Preliminary experimental work suggests that the model is on the right track, though the range of expression types could be expanded beyond modals, by taking into account a possible difference between the linguistic expression of necessity and possibility. Further analysis of our experimental data should shed further light on the differences between

linguistic expressions of (un)certainty and their interpretation, as well as the relationship between adverbials and epistemic modals. In addition, subjective certainty judgements can be used to empirically validate our temporal reasoning model, by comparing the model prediction of certainty of a temporal relation from the certainty of its component events, to the actual values rated by subjects. Finally, our experiment also needs to be extended to more specialised (and more fault-critical) cases, such as that of the NICU.

Acknowledgements

Thanks to Jim Hunter, Ehud Reiter, Kees van Deemter and Neil McIntosh for helpful comments on this work. We also thank the referees for their comments. Part of this work was supported by UK EPSRC grants EP/D049520/1 and EP/D05057X/1.

References

1. Terenziani, P., Snodgrass, R.T., Bottrighi, A., Torchio, M., Molino, G.: Extending temporal databases to deal with telic/atelic medical data. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS (LNAI), vol. 3581, pp. 58–66. Springer, Heidelberg (2005)
2. Barnett, G., Famiglietti, K., Kim, R., Hoffer, E., Feldman, M.: DXplain on the internet. In: Proceedings of AMIA1998, pp. 607–611 (1998)
3. Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., Sripada, S.: From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. AI Communications 22, 153–186 (2009)
4. Stacey, M., McGregor, C.: Temporal abstraction in intelligent clinical data analysis: A survey. Artificial Intelligence in Medicine 39(1), 1–24 (2007)
5. Allen, J.: Maintaining knowledge about temporal intervals. Communications of the ACM 26(11), 832–843 (1983)
6. Dubois, D., Allel, H., Prade, H.: Fuzziness and uncertainty in temporal reasoning. Journal of Universal Computer Science 9(9), 1168–1194 (2003)
7. Ryabov, V., Trudel, A.: Probabilistic temporal interval networks. In: Proceedings of TIME 2004, pp. 64–67 (2004)
8. Badaloni, S., Giacomin, M.: The algebra IA^{fuz}: a framework for qualitative fuzzy temporal reasoning. Artificial Intelligence 170(10), 872–908 (2006)
9. Palma, J., Juareza, J.M., Camposa, M., Marina, R.: Fuzzy theory approach for temporal model-based diagnosis: An application to medical domains. Artificial Intelligence in Medicine 38(2), 197–218 (2006)
10. Lai, A.M., Parsons, S., Hripcsak, G.: Fuzzy temporal constraint networks for clinical information. In: Proceedings of AMIA 2008, pp. 374–378 (2008)
11. Dechter, R., Meiri, I., Pearl, J.: Temporal constraint networks. Artificial Intelligence 49(1-3), 61–95 (1991)
12. Vila, L., Godo, L.: On fuzzy temporal constraint networks. Mathware & soft computing 1(3), 315–334 (1994)
13. Raufaste, E., da Silva Neves, R., Mariné, E.: Testing the descriptive validity of possibility theory in human judgements of uncertainty. Artificial Intelligence 148, 197–218 (2003)
14. Kratzer, A.: What *must* and *can* must and can mean. Linguistics and Philosophy 1, 337–355 (1977)
15. Grice, H.: Logic and conversation. In: Cole, P., Morgan, J. (eds.) Syntax and Semantics: Speech Acts. Academic Press, London (1975)
16. Papafragou, A.: Epistemic modality and truth conditions. Lingua 116, 1688–1702 (2006)
17. Klabunde, R.: Lexical choice for modal expressions. In: Proceedings of ENLG 2007 (2007)