Le projet BabyTalk : génération de texte à partir de données hétérogènes pour la prise de décision en unité néonatale

François Portet (1,2), Albert Gatt (1), Jim Hunter (1), Ehud Reiter (1), Somayajulu Sripada (1)

(1) Department of Computing Science, University of Aberdeen, Écosse
portet@imag.fr, {a.gatt,j.hunter,e.reiter,yaji.sripada}@abdn.ac.uk
(2) LIG, UMR 5217, Université de Grenoble, France

Résumé Notre société génère une masse d'information toujours croissante, que ce soit en médecine, en météorologie, etc. La méthode la plus employée pour analyser ces données est de les résumer sous forme graphique. Cependant, il a été démontré qu'un résumé textuel est aussi un mode de présentation efficace. L'objectif du prototype BT-45, développé dans le cadre du projet Babytalk, est de générer des résumés de 45 minutes de signaux physiologiques continus et d'événements temporels discrets en unité néonatale de soins intensifs (NICU**).** L'article présente l'aspect génération de texte de ce prototype. Une expérimentation clinique a montré que les résumés humains améliorent la prise de décision par rapport à l'approche graphique, tandis que les textes de BT-45 donnent des résultats similaires à l'approche graphique. Une analyse a identifié certaines des limitations de BT-45 mais en dépit de celles-ci, notre travail montre qu'il est possible de produire automatiquement des résumés textuels efficaces de données complexes.

Abstract Nowadays large amount of data is produced every day in medicine, meteorology and other areas and the most common approach to analyse such data is to present it graphically. However, it has been shown that textual summarisation is also an effective approach. As part of the BabyTalk project, the prototype BT-45 was developed to generate summaries of 45 minutes of continuous physiological signals and discrete temporal events in a neonatal intensive care unit (NICU). The paper presents its architecture with an emphasis on its natural language generation part. A clinical experiment showed that human textual summaries led to better decision making than graphical presentation, whereas BT-45 texts led to similar results as visualisations. An analysis identified some of the reasons for the BT-45 texts inferiority, but, despite these deficiencies, our work shows that it is possible for computer systems to generate effective textual summaries of complex data.

Mots-clés : Traitement automatique des langues naturelles ; Génération de texte ; Analyse de données ; Unité de soins intensifs ; Systèmes d'aide à la décision

Keywords: Natural language processing; Natural language generation; Intelligent data analysis; Intensive care unit; Decision support systems

1 Introduction

En unité de soins intensifs, le personnel médical est amené à prendre des décisions importantes concernant le meilleur traitement à administrer à un patient, parfois sous la pression du temps. Les cliniciens disposent d'un grand nombre de données qui incluent les signaux physiologiques (p.ex. fréquence cardiaque, pression artérielle...), les résultats de laboratoire, les notes qui enregistrent les interventions précédentes, etc. En principe, ce foisonnement d'informations devrait permettre une prise de décision appropriée, mais le plus souvent la quantité de données est telle qu'il est difficile de trouver l'information pertinente rapidement. Le mode de présentation des données est donc crucial. En effet, cette masse d'informations est utile dans la mesure où elle est présentée d'une manière qui permet aux éléments pertinents d'être extraits rapidement. Actuellement, le mode prédominant est la représentation graphique, mais, alors que les systèmes graphiques sont très performants lorsqu'ils sont utilisés par des experts (Shahar *et al.*, 2006), ils ne sont pas toujours efficaces dans le cas d'utilisateurs novices (dans notre cas, des infirmières ou des médecins débutants) devant prendre des décisions en quelques minutes. Une autre façon d'aider l'exploitation des données est de créer un système expert qui recommande des actions au personnel médical. Cependant, peu de ces systèmes ont été intégrés dans la pratique médicale car leurs recommandations sont souvent ignorées même lorsque les utilisateurs reconnaissent leur pertinence (Puppe *et al.*, 2008). Nous pensons qu'une manière alternative d'exploiter ces données pour l'aide à la décision est d'utiliser des systèmes à base de connaissances afin d'extraire les informations pertinentes et les présenter à l'utilisateur sous forme de résumé textuel produit automatiquement par des techniques de génération de texte (*Natural Language Generation – NLG*). Succinctement, il s'agit de trouver un compromis entre la présentation des données brutes (visualisation classique) et la recommandation d'actions (approche des systèmes experts). Notre but est donc de fournir aux cliniciens un résumé clair qui présente l'information principale pour faciliter la prise de décision, laissant cette dernière tâche entièrement à leur jugement. Ce travail a été initié au sein du projet britannique BabyTalk (Portet *et al*., 2009) dont le but est de générer des résumés textuels de données de nouveau-nés en unité néonatale de soins intensifs. Ce projet, ainsi que les données manipulées sont présentés en section 2. Le développement d'un prototype appelé BT-45 (BabyTalk 45 minutes) est ensuite résumé en section 3 et son évaluation est présentée en section 4. La section 5 termine l'article par une discussion des résultats et l'identification de certains défis pour la génération de langage naturel. Cet article est une version condensée de Portet *et al*. (2009).

2 Le projet BabyTalk

BabyTalk est une collaboration entre le NICU (*Neonatal Intensive Care Unit*) de la *Royal Infirmary of Edinburgh* et les universités d'Aberdeen et d'Édimbourg. L'objectif du projet est de concevoir des systèmes produisant des résumés textuels de données médicales à différentes échelles temporelles (minutes, heures, journées), pour différentes utilisations (aide à la décision, résumé pour l'équipe prenant le relais) et adaptés à différents utilisateurs (infirmières confirmées, médecins débutants, parents). Ces systèmes sont : 1) BT-Nurse (12 heures de données à destination des infirmières prenant leur service pour résumer ce qui s'est produit durant le service précédent) ; 2) BT-Doc (plusieurs heures de données, sur demande, pour aider les médecins débutants à prendre des décisions) ; 3) BT-Parent (production de résumés rassurants adaptés à l'état émotif des parents), une extension du système BabyLink (Freer *et al.*, 2005) qui est actuellement utilisé à Édimbourg ; et 4) BT-Clan (résumé à

destination des amis et de la famille, pour les encourager à offrir un soutien approprié aux parents). Avant d'atteindre ces objectifs, le prototype **BT-45 (BabyTalk 45 minutes)** a été conçu pour tester la faisabilité de l'approche – qui combine des techniques de traitement intelligent du signal et de génération de textes – sur de courtes périodes de données. La motivation principale pour la conception de BT-45 vient de l'étude de Law *et al.* (2005), qui ont constaté que les cliniciens en NICU prennent de meilleures décisions lorsqu'ils utilisent des données présentées sous forme textuelle (résumés uniquement descriptifs rédigés par des experts) plutôt que sous forme graphique, bien qu'ils préfèrent cette dernière forme de présentation (sûrement parce qu'ils y sont habitués). Notre première intention était d'essayer de reproduire ces résultats par le même type d'évaluation mais en présentant les données sous forme : **H**) textuelle rédigée par des experts humains, **C**) textuelle générée automatiquement, et **G**) graphique. Puisque l'étude de Law *et al.* (2005) a utilisé des scénarios de 45 minutes de données, BT-45 a été conçu pour produire des résumés de cette longueur.
Les données d'entrées de BT-45 ont été acquises durant le projet NEONATE (Hunter *et al.*, 2003) et représentent plus de 400 heures d'enregistrement de données de 42 nouveau-nés soignés dans la NICU de la *Royal Infirmary of Edinburgh*. Elles sont de deux sortes : 1) les événements discrets (annotations d'intervalles de temps) tels que les résultats de laboratoire, les activités du personnel, les équipements utilisés et leur paramétrage, les prescriptions, etc. et 2) les données multivoies continues de signaux physiologiques (fréquence cardiaque *HR*, pressions de l'oxygène *TcPO2* et de l'anhydride carbonique *TcPCO2* dans le sang, saturation en oxygène *SpO2*, températures périphérique *T2* et centrale *T1* et tension artérielle moyenne *mean BM*). La Figure 1-G) donne un aperçu des données d'entrée du système telles que les cliniciens les perçoivent habituellement. La seule différence avec les systèmes réels est que les données discrètes de NEONATE ont été saisies précisément par une infirmière employée à cet effet. En plus des données cliniques, nos collègues cliniciens à Édimbourg ont écrit 23 résumés, décrivant des périodes de temps entre 30 et 50 minutes, qui ont complété les 18 résumés écrits pour le projet NEONATE (Hunter *et al.*, 2003). Ces résumés ont été employés comme données de développement pour BT-45. Vingt-six autres résumés ont été écrits pour évaluer la prise de décision clinique à partir de résumés écrits par des experts ou générés par BT-45. La Figure 1-H) montre un texte, écrit par un expert, correspondant aux données de la Figure 1-G).

3 Architecture du prototype BT-45

L'architecture du prototype est illustrée Figure 2. BT-45 crée un résumé des données cliniques en quatre étapes principales. Les signaux physiologiques et les annotations sont traités par l'étape d'*analyse des signaux* (1) pour extraire les caractéristiques principales des signaux (événements médicaux, artéfacts et tendances). Le module d'*interprétation* (2) abstrait les événements (p.ex. abstraction d'événements consécutifs en séquence) et interprète les données pour découvrir des relations entre événements (p.ex. A cause B). À partir du grand nombre d'événements et de relations produites, l'étape de *planification du document* (3) choisit les plus importants et les agrège dans un arbre d'événements. Enfin, le module de *micro-planification et réalisation* (4) traduit cet arbre en texte. Tous les termes employés pour décrire les événements sont liés par une ontologie des concepts du domaine NICU. Les sous-sections suivantes détaillent les modules de BT-45 en mettant l'accent sur la génération de texte.

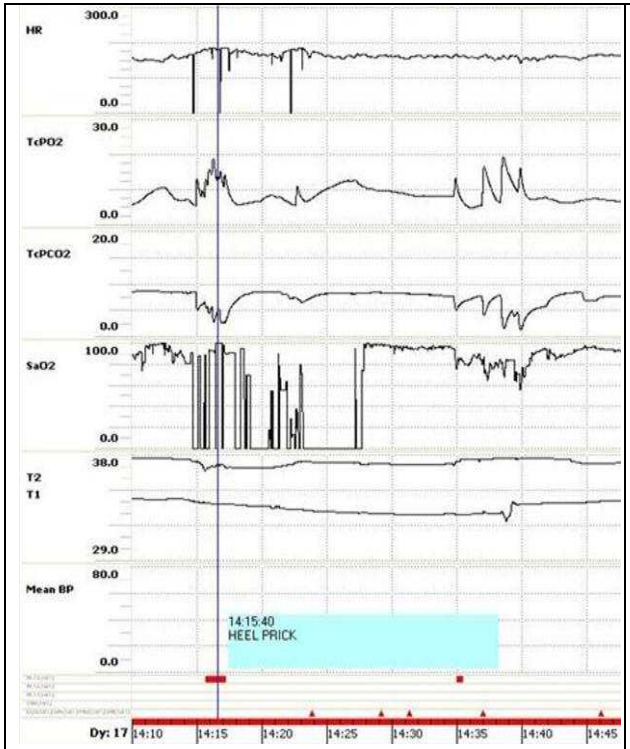	To start with, the HR=152; TcPO2= 6; TcPCO2= 8.6 and SaO2= 92%. T1= 37.7º and T2= 34.3ºC. Initially there is an increase in oxygenation: TcPO2 rises from 6 to 10.6 SaO2 rises from 89 to 98%. HR and TcPCO2 are static. At 14:15 hours a heel prick is done. The HR increases at this point and for 7 minutes from the start of this procedure there is a lot of artifact in the oxygen saturation trace. Also there are upward spikes in the TcPO2 (to 17), each corresponding with downward spikes in the TcPCO2 (to 2.6). At 14:23 the FiO2 is changed to 45% and the TcPO2 rises from 6 to 12. At 14:28 the saturation probe is reapplied, the oxygen saturation is 100%, and the HR settles back to baseline of 155. The FiO2 is reduced twice and at 14:31 is 25%. At 14:34 hours the baby is examined. For the following 5 minutes there is another series of spikes in the transcutaneous parameters, with TcPO2 spikes up to 19 and corresponding TcPCO2 spikes down to 1.4, while the saturations fall to 74%. The FiO2 is increased to 32%. HR remains stable at 145-160. By 14:38 the toe-core gap has widened to >4º with a T1 of 37.7ºC. Both temperature probes are re-applied at 14:41. Shortly after this the oxygen saturation settles to 94-95% and at 14:46 the FiO2 is reduced to 28%. At 14.48 hours the oxygen saturation falls to 83%, TcPO2 is 6.6 and TcPCO2 is 7.6. HR remains steady around 160. While the oxygen saturation remains at 83% the TcPO2 falls to 5 and the TcPCO2 to 6.6. The FiO2 is increased to 30%. By 14:50 T1 is 37.5º and T2 is 34.2ºC.	You saw the baby between 14:10 and 14:50. Heart Rate (HR) = 159. Core Temperature (T1) = 37.7. Peripheral Temperature (T2) = 34.3. Transcutaneous Oxygen (TcPO2) = 5.8. Transcutaneous CO2 (TcPCO2) = 8.5. Oxygen Saturation (SaO2) = 89. Over the next 30 minutes T1 gradually increased to 37.3. By 14:27 there had been 2 successive desaturations down to 56. As a result, Fraction of Inspired Oxygen (FIO2) was set to 45%. Over the next 20 minutes T2 decreased to 32.9. A heel prick was taken. Previously the spo2 sensor had been re-sited. At 14:31 FIO2 was lowered to 25%. Previously TcPO2 had decreased to 8.4. Over the next 20 minutes HR decreased to 153. By 14:40 there had been 2 successive desaturations down to 68. Previously FIO2 had been raised to 32%. TcPO2 decreased to 5.0. T2 had suddenly increased to 33.9. Previously the spo2 sensor had been re-sited. The temperature sensor was re-sited.
G)	***H)***	***C)***

Figure 1: **G)** Présentation **G**raphique de données NICU. Les voies physiologiques de haut en bas sont HR, TcPO2, TcPCO2, SaO2, T1 & T2, et Mean BP (non connecté durant la période d'enregistrement). Les carrés rouges en bas représentent les événements annotés, l'utilisateur accède à de l'information supplémentaire en cliquant dessus (p.ex., un *heel prick* - acquisition d'échantillon sanguin au talon - a été enregistré à 14h15). **H)** Résumé textuel correspondant écrit par un expert **H**umain. **C)** Résumé textuel correspondant généré par BT-45 (**C**omputer)

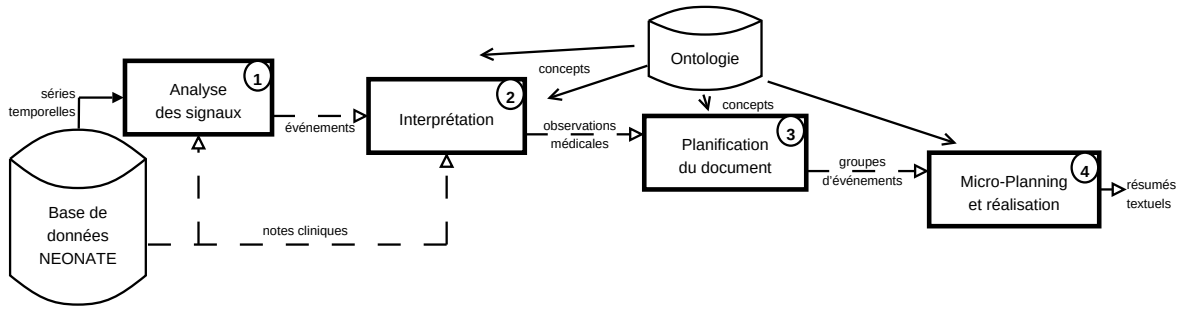

Figure 2: Architecture du prototype BT-45.

3.1 Analyse des données et représentation dans l'ontologie du domaine

Dans BT-45, la connaissance du domaine est centralisée dans une ontologie. En plus d'avoir fourni un vocabulaire conceptuel commun entre les modules, cette ontologie a également servi au raisonnement. Plutôt que d'utiliser des ressources médicales telles que SNOMED-CT ou UMLS trop difficiles à intégrer dans un système dédié (UMLS couvre plus de 1.5 million de concepts), nous avons développé cette ontologie à partir d'un lexique créé durant nos précédents projets en incluant notamment des informations linguistiques. La version finale représente environ 550 concepts séparés en deux branches principales : EVENT et ENTITY. ENTITY englobe les objets du domaine, tels que NURSE, VENTILATOR (appareil d'assistance respiratoire), MEDICATION, etc. EVENT englobe les activités et les événements qui impliquent des entités. Tous les événements ont une date de début, une date de fin, et une valeur d'importance. Cette dernière représente l'importance médicale d'un événement qui peut être fixe ou calculée par BT-45. Les sous-classes de EVENT incluent INTERVENTION (p.ex., injection de médicament), OBSERVATION (p.ex., bébé est agité), etc. Étant donné que l'ontologie est utilisée pour représenter la connaissance et pour soutenir le traitement linguistique, les événements ont des attributs qui spécifient leurs participants. Durant la lexicalisation, ces participants sont reliés aux rôles thématiques d'une structure événementielle utilisée pour exprimer un événement. Par exemple, INSERT_CHEST_DRAIN a des champs qui spécifient l'agent (la personne qui insère le drain), le bénéficiaire (la personne à qui le drain est destiné), et le thème (l'objet qui a été utilisé; ici le drain).

Les instances des classes de l'ontologie sont créées lors de la lecture de la base de données ou inférées par les modules d'analyse des signaux ou d'interprétation des données. Les signaux physiologiques contiennent beaucoup d'informations sur l'état du patient qui doivent être extraites (p.ex. les bradycardies – périodes de ralentissement du rythme cardiaque). Le but de l'analyse des signaux est donc de détecter les motifs représentant des événements d'intérêt sur les signaux physiologiques, de les classifier et d'estimer leur importance. Toutes les instances sont ensuite analysées par le module d'interprétation des données qui est un prérequis pour la synthèse des données. En effet, reporter les événements pris isolement (p.ex. chaque pic) ne réduirait pas la charge d'informations tandis que l'abstraction de ceux-ci dans une description de plus haut niveau (p.ex., une séquence de pics) est déjà une compression d'information. Par ailleurs, les associations entre couples d'événements (p.ex., A cause B) doivent être inférées pour faciliter la compréhension. Ceci est réalisé en utilisant plus d'une centaine de règles et métarègles expertes. Le lecteur qui souhaiterait plus de détails sur ces parties pourra se reporter à l'article de Portet *et al.* (2009).

3.2 Planification du document

Ce module décide, à partir des événements et des liens produits par les modules en amont, quels sont les événements (c.-à-d. « unités d'informations » ou « messages ») devant être communiqués dans le texte. La planification répartit les messages dans des paragraphes et détermine l'ordre des événements dans chaque paragraphe. Le plan résultant est une représentation abstraite d'un document sous forme d'un arbre dont les noeuds sont des événements (messages) ou des informations de structure du document (paragraphes), et dont les arcs sont annotés avec des relations rhétoriques. La Figure 3 illustre la manière de fonctionner de la planification. Le noeud racine (*root*) a quatre fils : un noeud SEQUENCE (l'événement clé – *key event*), avec les événements qui lui sont liés (DESATURATION et FIO2) ; et les autres noeuds de groupe construits en fonction de leur catégorie (système respiratoire, thermorégulation et autre...). L'algorithme, inspiré de Hallett *et al.* (2006), identifie un nombre restreint d'événements clés (basé sur leur importance) et crée un paragraphe pour chacun. Les événements clés sont mentionnés d'abord, suivis des événements qui leur sont explicitement liés, suivis de ceux se produisant au même moment. Les paragraphes sont ordonnés par heure de début de leur événement clé respectif. L'originalité de cette approche réside dans l'algorithme de construction autour d'événements clés et par le fait qu'un paragraphe est traité comme un type primitif. Dans la plupart des systèmes NLG, les paragraphes ont tendance à suivre des modèles très stricts (par exemple, un paragraphe au sujet des médicaments) ou à être traités comme un phénomène d'agrégation. Dans BT-45, l'algorithme produit dynamiquement des paragraphes de longueur et de contenu variable.

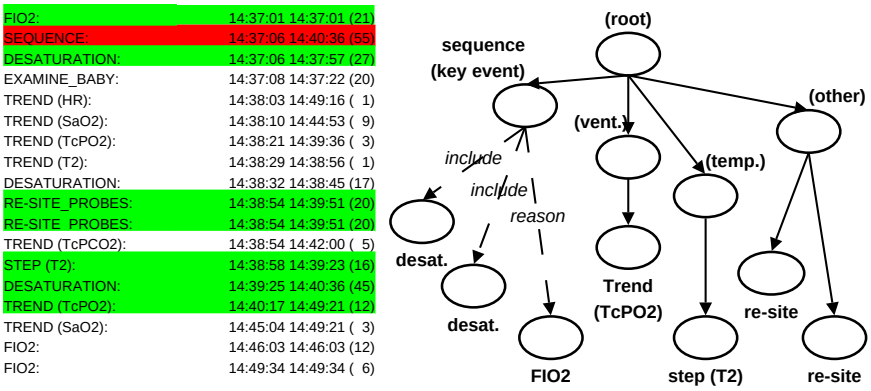

Figure 3 **:** Liste d'événements pour la période de 14h37 à 14h50 (à gauche), dont les événements grisés ont été sélectionnés pour constituer une partie du plan du document (à droite). Les noeuds de groupement, entre parenthèses, ne correspondent pas à un événement.

3.3 Micro-planification et réalisation

Ce module regroupe le micro-planning et de réalisation qui sont souvent considérées séparées en NLG (Reiter et Dale, 2000). Le microplanning ajoute le contenu linguistique au plan de document, créant des représentations sémantiques qui sont ensuite traduites en structures linguistiques (syntactiques), pour être finalement linéarisées comme texte. L'architecture interne du microplanner BT-45 est donnée **Figure 4**. Les différentes parties du microplanner sont illustrées en se référant au dernier paragraphe du texte de la Figure 1-C).

La lexicalisation relie les événements du plan à des structures événementielles (*Event-Frame*) consistant en un verbe (prédicat) et une spécification de ses arguments sémantiques (rôle thématique) tels que AGENT ou THEME. La lexicalisation associe les EVENT à des

prédicats en se basant sur leur classe ontologique et en utilisant des règles expertes. Ce procédé est également soutenu par le lexique Verbnet (Kipper *et al.*, 2000) modifié pour inclure des classes spécifiques au domaine. Verbnet groupe les verbes par classes selon leurs rôles thématiques. Par exemple, dans le cas de l'événement où le paramètre du ventilateur FIO2 est augmenté à 32%, cet événement appartient à une sous-classe de VENTILATOR_SETTING. Cette classe possède un attribut de direction qui, dans ce cas-ci, prend la valeur *increase* de sorte que l'exemple correspond au modèle suivant (au format LISP) : `(event-verb-mapping (event-class VENTILATOR_SETTING) (verb-class intentional_value_setting) (direction increase) (verb raise))`. Le verbe appartient à la classe *intentional_value_setting* du lexique et hérite de trois rôles thématiques, AGENT (l'entité qui modifie FIO2), THEME (l'entité qui est modifiée, ici FIO2) et VALUE (ici, 32%). La lexicalisation est généralisée pour traiter des séquences d'événements, en groupant ces derniers dans une seule structure événementielle. L'*Event Linking* cherche à rendre explicite les relations entre les événements du plan. Les relations temporelles sont exprimées en utilisant des adverbiaux temporels. D'autres types de liens, tels que les relations causales, sont traités par heuristique pour choisir la manière la plus adéquate de les exprimer. Par exemple, si la cible d'un lien causal est une structure événementielle réalisée comme clause non existentielle et déclarative, celle-ci sera réalisée comme une clause séparée telle que *as a result*. Par contre, les clauses existentielles (p.ex., *there was a bradycardia*) sont réalisées en tant que clauses subordonnées (p.ex., *The baby was given morphine, causing a bradycardia*). Suite à ces opérations une structure événementielle contient un certain nombre de rôles thématiques qui se réfèrent à des entités du domaine, pour lesquelles des expressions référentielles (RE) doivent être construites. Le module GRE (*Generation of Referring Expressions*) manipule quatre types de RE : 1) les *entités nommées* ; 2) les *mass term* qui se rapporte à des substances ou des génériques (p.ex. MORPHINE dans *50mg of morphine*) ; 3) les groupes nominaux (GN) définis et indéfinis, qui sont identifiés par leurs propriétés dans l'ontologie (p.ex. *the baby* est un GN défini unique dans le discours alors que *an IV line* est toujours indéfini) ; et 4) les références anaphoriques, pour lesquelles un algorithme basé sur la prépondérance (Krahmer et Theune 2002) est employé pour déterminer si des entités doivent être mentionnées par des pronoms.

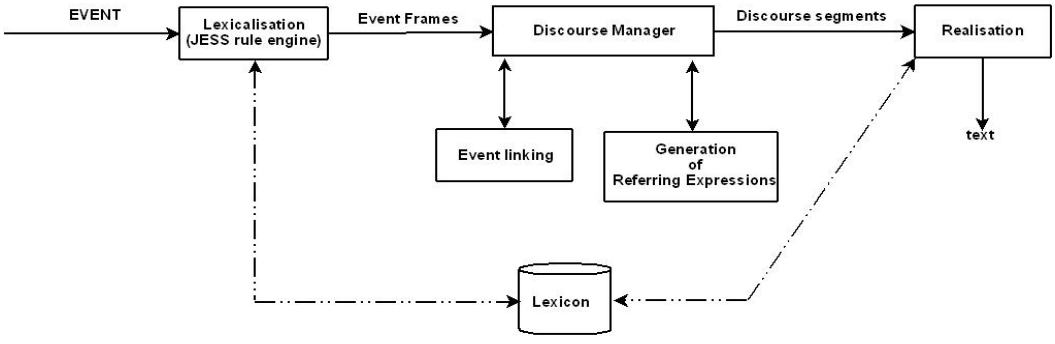

Figure 4 : Architecture du module de micro-planification de BT-45

Un des points les plus délicats pour la micro-planification est l'expression du temps qui est gérée par le gestionnaire du discours. Le lecteur doit pouvoir reconstruire, à partir du texte, l'ordre dans lequel les événements se sont produits. Cependant, l'ordre narratif de BT-45 n'est pas isomorphe à l'ordre temporel. Ceci est dû à la stratégie d'ordonnancement du planificateur (basée sur l'importance) que le microplanner essaye de respecter. Malgré la quantité substantielle de publications sur la représentation temporelle dans le traitement du langage naturel, ce problème n'a suscité que très peu d'attention du point de vue de la génération.

L'événement clé qui forme la racine de chaque paragraphe est toujours exprimé avec une mention explicite de son heure de début, de sorte que chaque paragraphe commence par une référence temporelle claire. Des temps et des adverbiaux temporels sont ensuite employés pour indiquer l'ordre temporel relatif des événements mentionnés après l'événement clé. Les temps sont calculés en utilisant le modèle proposé par Reichenbach (1947/1966). Le temps utilisé dépend du moment d'occurrence de l'événement (E), du moment de l'énonciation, et du moment de référence (R). Stylistiquement, ceci distingue les textes BT-45 écrits au passé des textes humains où le présent de narration est employé. L'ordre relatif de R et de E pour un ensemble d'événements détermine l'utilisation d'un *past perfect* ou d'un *simple past*. Par exemple, *T2 had suddenly increased to 33.9* indique que son moment d'occurrence précède son moment de référence (E < R). Dans ce cas-ci, son moment de référence est le moment d'occurrence de l'événement mentionné précédemment qui a en fait commencé après lui. La phrase *The temperature sensor was re-sited* a aussi l'événement mentionné précédemment comme moment de référence mais, comme les deux événements se sont produits au même moment (R = E), cette phrase est exprimée au *simple past*.

4 Évaluation

Pour évaluer BT-45, il a été demandé à 35 infirmières et médecins de la *Royal Infirmary of Edinburgh* de prendre une décision à partir de 45 minutes d'informations sur des nouveau-nés (qu'ils ne connaissaient pas) présentées sous forme : graphique **(G)** ; textuelle rédigée par nos experts humains **(H)** ; ou textuelle générée par BT-45 **(C).** 24 scénarios ont été choisis pour lesquels 18 actions (appropriées, neutres ou inappropriées) pouvaient être choisies. Ces 24 scénarios, sous trois conditions, ont été distribués en utilisant une stratégie de carré latin pour éviter qu'un sujet soit confronté plusieurs fois au même scénario. Les textes **H** ont été écrits pour être uniquement descriptifs et pour éviter toutes indications thérapeutiques (p.ex., *HR is* ***normal*** ou *arterial pressure is* ***worrying***). Les sujets ont été répartis en quatre groupes : docteurs seniors SD (n=9) et juniors JD (n=9), infirmières seniors SN (n=9) et juniors JN (n=8). Chaque scénario devait être analysé en moins de 3 minutes afin d'imposer une certaine pression. Les sujets n'ont pas été invités à donner leur avis mais tout avis spontané a été enregistré anonymement. Avant l'évaluation, une initiation au logiciel de test a été effectuée. Aucun des développeurs de BT-45 n'a eu accès aux scénarios avant l'expérimentation. La longueur moyenne des 26 (24 scénarios + 2 scénarios d'apprentissage) textes **H** était de 135 mots (σ = 79) et de 119 mots (σ = 36) pour les textes **C** mais ils n'ont pas une longueur significativement différente selon le test de Wilcoxon (z = -.588, p = .493). Bien que les textes **H** soient plus longs ils partagent la même tendance (c.-à-d., quand les textes **H** sont plus longs, les textes **C** le sont aussi). Le score de chaque sujet pour chaque scénario a été calculé en soustrayant la proportion d'actions choisies inadéquates de la proportion d'actions choisies appropriées. Le score global pour **G** était 0.33 (σ = 0.14), pour **H** 0.39 (σ = 0.11) et pour **C** 0.34 (σ = 0.14). Un test ANOVA 3 (conditions) x 4 (groupes) a montré un effet de condition approchant la significativité ($F(2, 31) = 2.939$, $p = 0.06$) mais aucun effet de groupe, et aucune interaction. Trois analyses ANOVA ont montré que **G** et **H** ($F(1, 31) = 4.975$, $p < 0.05$) et que **C** et **H** ($F(1.31) = 5.266$, $p < 0.05$) donnent des résultats significativement différents alors que les résultats de **G** et **C** ne le sont pas (van der Meulen *et al.*, 2009).

5 Discussion

Le résultat le plus important de cette étude est sûrement d'avoir montré qu'il est possible de générer des résumés de données cliniques complexes pouvant aider efficacement la prise de

décision. En effet, les décisions sous conditions **C** sont au moins aussi efficaces que sous condition **G** (et même bien meilleures dans certains cas) alors que G est une présentation quotidiennement utilisée par le personnel et que BT-45 n'est le résultat que d'un an de développement. Un autre résultat important est que les textes peuvent être une aide à la décision très efficace comme le prouvent les résultats de **H** significativement meilleurs que **C** et **G**. Ceci conforte les résultats de Law *et al* (2005) et encourage le développement de nouvelles technologies de génération de textes à partir d'ensembles de données temporelles hétérogènes. Afin de nous aider à comprendre quels aspects des textes **C** les ont rendus moins efficaces que les textes **H**, nous avons analysé les résultats en profondeur. Cette analyse nous a permis d'identifier les principales limites du système à tous les niveaux de traitement. Nous résumons ici les principaux problèmes liés à la génération de texte.

Le principal aspect de BT-45 qui a été critiqué est la description trop morcelée des signaux physiologiques. Par exemple, le texte de la Figure 1-C) commence avec *T1 is 37.7*, et *TcPO2 is 5.8*, et les prochaines références sont *T1* ***increased*** *to 37.3*, et *TcPO2* ***decreased*** *to 8.4.* Ce problème est dû à la stratégie de planification *bottom-up* de BT-45. Sur la Figure 1-G) le signal TcPO2 s'est élevé, entre 14.15 et 14.17, à une valeur de crête de 20, avant de diminuer à 8.4. Les règles expertes ont assigné plus d'importance à la chute de TcPO2 qu'à l'élévation précédente, ce qui a provoqué la sélection unique de cette première par le planificateur. Nous appelons ce problème *continuity* qui fait référence à l'expression employée par des réalisateurs de film lorsqu'ils cherchent à s'assurer que les scènes voisines dans un film « collent » les unes avec les autres lors du montage. Certains textes humains ont également des problèmes de continuité, mais aucun des sujets ne s'en est plaint. Ceci suggère que certaines violations de continuité sont plus problématiques que d'autres et il se peut que cela soit lié à la proximité des événements dans le temps et dans l'espace du texte.
Un autre problème est la mauvaise communication des informations temporelles. L'une des difficultés est de choisir les références temporelles pour un événement. Par exemple dans un scénario, BT-45 a généré : *After 6 attempts, at 14:17 a peripheral venous line was inserted successfully.* En fait, 14 :17 correspond à la date de la première tentative or les lecteurs l'interprètent comme la date de succès de l'opération. Ceci est d'autant plus gênant que cette date, pour un événement aussi long, peut servir de date de référence pour les événements suivants du texte. Le problème est que quand BT-45 décrit de longs événements, il ne prend pas en compte le type de l'événement. Dans cet exemple, la notion de succès est cruciale. En effet, si tous les essais avaient été infructueux cette date serait acceptable. BT-45 a donc besoin d'un meilleur modèle pour communiquer le temps et pour lier l'expression linguistique à la sémantique des événements. Nous avons également besoin de meilleures méthodes pour communiquer les relations temporelles entre les événements dans les cas où ils ne sont pas énumérés dans l'ordre chronologique.
D'après deux analystes du discours de l'université d'Édimbourg (Andrew McKinlay et Chris McVittie), la supériorité des textes humains peut être expliquée par leur structure narrative, bien meilleure que celle des textes BT-45. Ces analystes emploient le terme narratif dans le sens de Labov (1971), c'est-à-dire des expériences écrites comme des récits et qui vont au-delà d'une description factuelle d'événements et qui incluent les informations qui aident des auditeurs à rendre réaliste ce qui s'est produit (p.ex. : résumés, évaluatifs, corrélatifs, ou explicatives). En fait, une grande partie des limitations du système (continuité, information relative, granularité temporelle et vue d'ensemble) sont des aspects de la génération narrative. Ceci est d'autant plus intéressant que nos collaborateurs cliniciens à Édimbourg sont persuadés de la valeur des informations sur les nouveau-nés rapportées sous forme de récits. En effet, les systèmes de gestion de patients actuels, où la saisie d'informations est réalisée à travers des formulaires, ont supprimé les notes dites « libres » qui contenaient des erreurs (orthographiques, typographiques, etc.) mais dont la structure narrative contenait des

informations précieuses (relations causales et explications). Or, les cliniciens débutants apprenaient beaucoup grâce à ces notes qui rendaient les liens de cause/conséquence entre les événements explicites.

Toutes ces limitations sont prises en compte pour le développement des prochains systèmes du projet BabyTalk destinés à un éventail plus large d'utilisateurs. Nous pensons que cette technologie de *données-vers-texte* sera bientôt mature et pourra être employée pour aider des personnes à comprendre de grands ensembles de données, en médecine, en météorologie et dans beaucoup d'autres secteurs.

Remerciements

Les auteurs remercient les membres du projet BabyTalk et les cliniciens ayant participé à l'évaluation pour leur aide. Ce travail est soutenu par *Engineering and Physical Sciences Research Council* (EPSRC) à travers les financements EP/D049520/1 et EP/D05057X/1.

Références

FREER Y., LYON A., STENSON B., COYLE C. (2005). BabyLink – improving communication among clinicians and with parents with babies in intensive care. *Br J Healthcare Comput Inf Manage*, 22(2), 34-36.

HUNTER J., FERGUSON L., FREER Y., EWING G., LOGIE R., MCCUE P., MCINTOSH N. (2003) The NEONATE Database. *Joint AIMDM and IDAMAP Workshop*, 21-24.

HALLETT C., POWER R., SCOTT D. (2006). Summarisation and visualisation of e-Health data repositories. *UK E-Science All-Hands Meeting*, Nottingham, UK.

KIPPER K., DANG H.T., PALMER M. (2000). Class-based construction of a verb lexicon. *17th National Conference on Artificial Intelligence*, Austin, Texas, 691–696.

KRAHMER E., THEUNE M. (2002). Efficient context-sensitive generation of referring expressions. *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, Stanford, CA.

LABOV W. (1971) Language in the Inner City. University of Pennsylvania Press, Pennsylvania.

LAW A.S., FREER Y., HUNTER J., LOGIE R.H., MCINTOSH N., QUINN J. (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit, *J Clin Monit Comput*, 19(3), 183-194.

PORTET F., REITER E., GATT A., HUNTER J., SRIPADA S., FREER Y., SYKES C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artif Intell*, 173(7-8), 789-816.

PUPPE F., ATZMÜLLER M., BUSCHER G., HÜTTIG M., LUEHRS H., BUSCHER H.-P. (2008). Application and Evaluation of a Medical Knowledge System in Sonography (SONOCONSULT). *ECAI*, 683-687

REICHENBACH H. (1947/1966). *Elements of Symbolic Logic*, Macmillan, New York.

REITER E., DALE R. (2000). *Building Natural Language Generation Systems*, Cambridge University Press, Cambridge.

SHAHAR Y., GOREN-BAR D., BOAZ D., TAHAN G. (2006). Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions, *Artif Intell Med*, 38(2), 115-135.

VAN DER MEULEN, M., LOGIE R.H., FREER Y., SYKES C., MCINTOSH N., HUNTER J. (2009). When a graph is poorer than 100 words: A comparison of computerised Natural Language Generation, human generated descriptions and graphical displays in neonatal intensive care, *Appl Cogn Psychol*.