Génération automatique de résumés textuels de données multimodales pour l'aide à la décision en unité néonatale

François Portet

Laboratoire d'Informatique de Grenoble – Équipe GETALP
francois.portet@imag.fr

25 mars 2011

Laboratoire d'Informatique de Grenoble (LIG)

- ▶ Créé en 2007
- ▶ 500 chercheurs
- ▶ Université Joseph Fourier, Grenoble INP, Université Pierre-Mendès-France, CNRS, INRIA Grenoble Rhône-Alpes
- ▶ 4 thèmes scientifiques majeurs : infrastructures (réseaux, informatique ambiante et durable), logiciel (GL), Interaction (perception, action et dialogue), connaissances (apprentissage, agents, ontology-web).

GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole)

- ▶ 12 permanents
- ▶ 3 prof/HDR : H. Blanchon, C. Boitet & L. Besacier
- ▶ six thèmes de recherche principaux :
  1. Traduction Automatique (TA) et Automatisée (TAO)
  2. Traitement Automatique des Langues (TALN) et plates-formes associées
  3. Collecte et construction de ressources linguistiques
  4. Multilinguisme dans les systèmes d'information
  5. Reconnaissance automatique de la parole, des locuteurs, des sons et des dialectes
  6. Analyse sonore et interaction dans les environnements perceptifs

Projets de l'équipe

- ANR OMNIA : Indexation, catégorisatation et organisation de documents multimédia
- ANR PORT-MEDIA : Robustness and Multilingual multidomain portability of spoken language understanding systems
- ANR QCOMPERE : Quaero Consortium for Multimodal Person Recognition
- ANR VideoSense : Automatic video tagging by high level concepts, including static concepts (e.g. object, scene, people, etc.), events, and emotions, while targeting two applications, namely video recommendation and ads monetization.
- ANR SWEETHOME : Système Domotique d'Assistance au Domicile (`sweet-home.imag.fr`)

Data-to-Text (E. Reiter)

GAT
Production de textes à partir de données non linguistiques
[Danlos and Roussarie, 2000]

Data-to-Text
Grande quantité de données hétérogènes (de type numérique)

Deux problèmes fondamentaux en GAT

1. Quoi dire ? (Génération profonde)
2. Comment le dire ? (Génération de surface)

Domaines connexes

1. Résumé automatique de textes (mais le "comment dire" est le plus souvent résolu)
2. Traduction automatique (mais le "quoi dire" est résolu)

Some Applications

- Weather report generation from Numerical Weather Prediction models (FOG, MultiMeteo, SumTime-Meteo).
- Suregen system used by physicians to create surgical reports [Hüske-Kraus, 2003].
- Tailored smoking cessation letters based on a user questionnaire [Reiter et al., 2003].
- Short stories based on a story plan [Callaway and Lester, 2002].

BabyTalk Outlines

Aim
Improving Information Flows in neonatal intensive care units (NICU) by Automatic generation of texts summarising baby's ICU data

Team
UK project by the BabyTalk Team : University of Aberdeen and Edinburgh (`www.csd.abdn.ac.uk/research/babytalk/`)

Scientific Aim & Challenges
Build a framework for summarisation as text of large heterogeneous multimodal medical data.

What for ?
- ▶ Better understanding of data
- ▶ Decision support
- ▶ Adaptation to different audiences

BabyTalk : For whom ?

Professionals don't like to be told what to do : summarisation may be a better way to support medical decision, but not only.

- ► Doc → medical decision.
- ► Nurse → medical decision and paper work.
- ► Family → wants to know what is going on.

Four systems

- ► BT-Nurse : 12 hours of data to serve as a shift summary
- ► BT-Doc : 10-12 hrs, support decision making by junior doctors
- ► BT-Family & BT-Clan : daily summary for the baby's family, adapted to the emotional state of the recipient.
- ► BT-45 : automatic summarisation of 45 minutes of NICU data

Motivations, why text ?

Graphical presentation in temporal Data intensive environments

+ Capacity of summarising a large amount of information
+ "Intuitive"

Can you interpret this ?
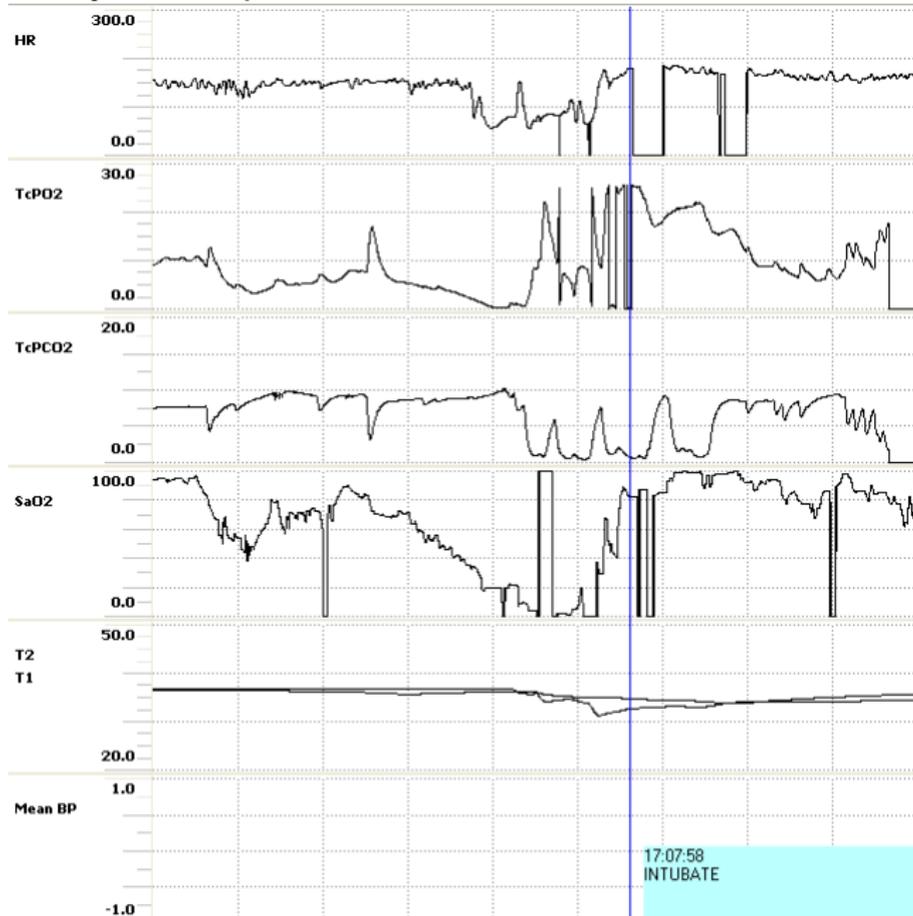
Motivations, why text ?

Graphical presentation in temporal Data intensive environments

- $+$ Capacity of summarising a large amount of information
- $+$ "Intuitive"
- - Granularity representation
- - Training
- - Human limits

Text summarisation

- - Might take time to read
- - Less details (not raw data)
- $+$ No training needed
- $+$ Capacity of representing different granularity
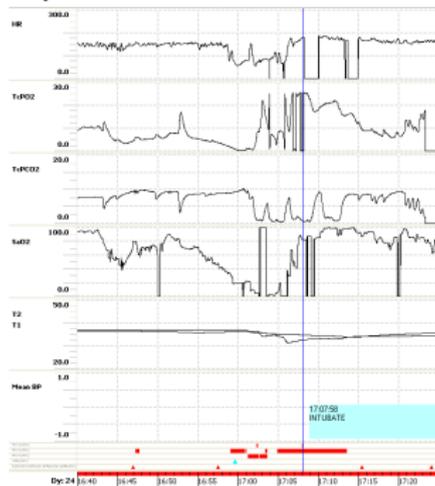- $+$ Filtering/Compression

Which one is the best : the graph vs. text experiment (Law et al. 2005) ?

Can we generate text automatically ?

Can we take advantage of both presentations ? (not investigated in BabyTalk)

Is it Better ?

input

target human text

"...The pCO2 continues to rise to 10.1. The baby is pale and unresponsive. ET suction is given, baby is turned and at 17 :02 the ETT is removed ; the baby is again given Neopuff ventilation. Baby is re-positioned and the NGT aspirated. By 17 :08 the baby is reintubated ; the oxygen saturation has increased to the 80s and the HR has risen to 176 the pO2 = 0.1 and pCO2 = 0.2, T1 is 32.7°C and T2 34.7°C.
At 17 :15 the FiO2 is reduced to 33% and the rate put back to 15 ..."

Is it Better ?

input

target human text

"...The pCO2 continues to rise to 10.1. The baby is pale and unresponsive. ET suction is given, baby is turned and at 17:02 the ETT is removed; the baby is again given Neopuff ventilation. Baby is re-positioned and the NGT aspirated. By 17:08 the baby is reintubated; the oxygen saturation has increased to the 80s and the HR has risen to 176 the pO2 = 0.1 and pCO2 = 0.2, T1 is 32.7°C and T2 34.7°C.
At 17:15 the FiO2 is reduced to 33% and the rate put back to 15 ..."

NICU environement
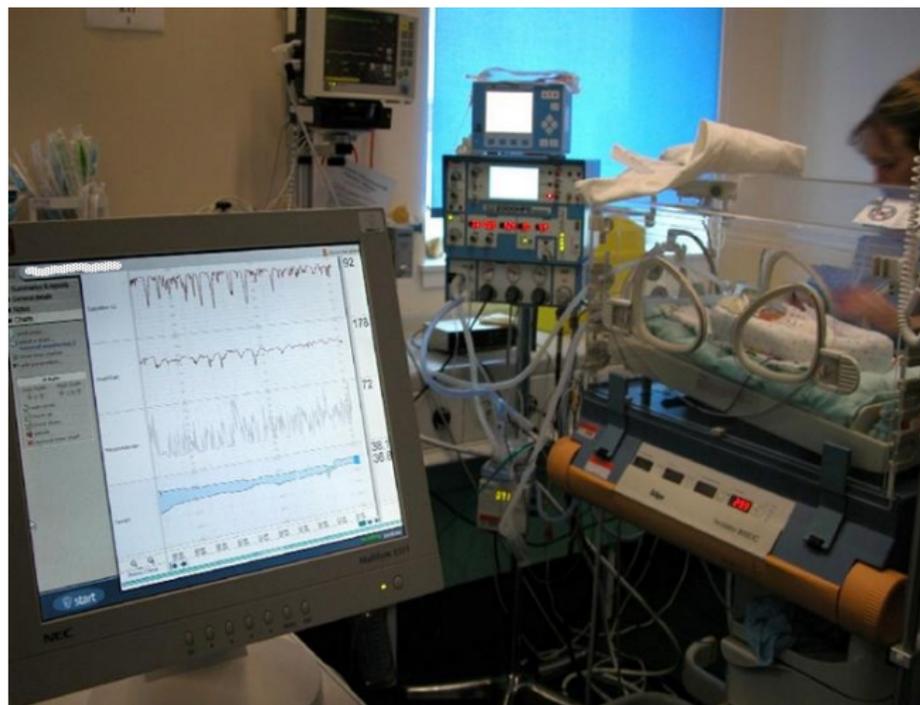
We talk about : Monitors, Nurses, Baby, Medication, Settings,
Signal, Feeding, . . .

Data available

Continuous
Monitor (1 Hz) : heart rate, blood pressures, O2, CO2, temperatures . . .(i.e. 86,400 samples/channel/patient/day)

Sporadic

- ▶ Equipment setting : ventilator, incubator . . .
- ▶ Investigation results : on-ward blood gases, laboratory . . .
- ▶ Medication : drug administration . . .
- ▶ Free text : notes, reports . . . (not used in BT-45)

Special to BT-45
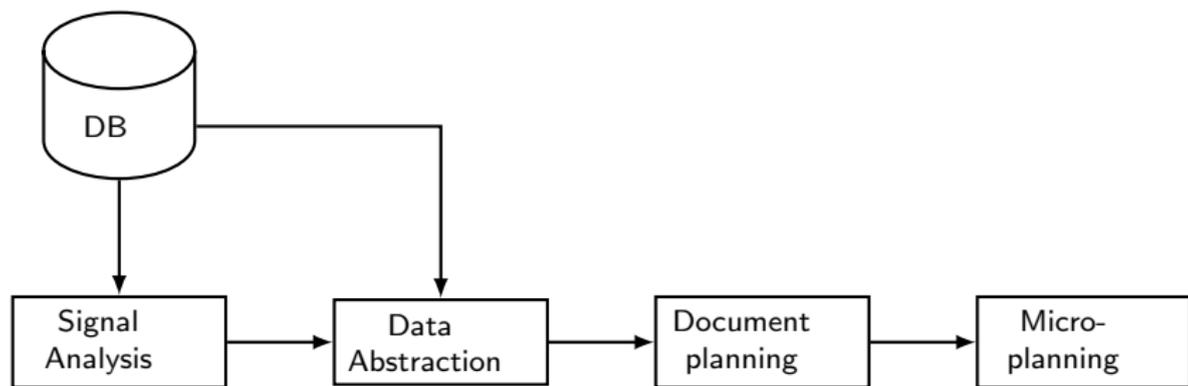Annotations of clinical events : baby is being handled, baby is intubated . . .

How to do ?

- ▶ What are the different steps ?
- ▶ How to represent and use domain knowledge ?
- ▶ How to extract relevant information from raw data ?
- ▶ How to select what to say ?
- ▶ How to write correct concise medical English ?
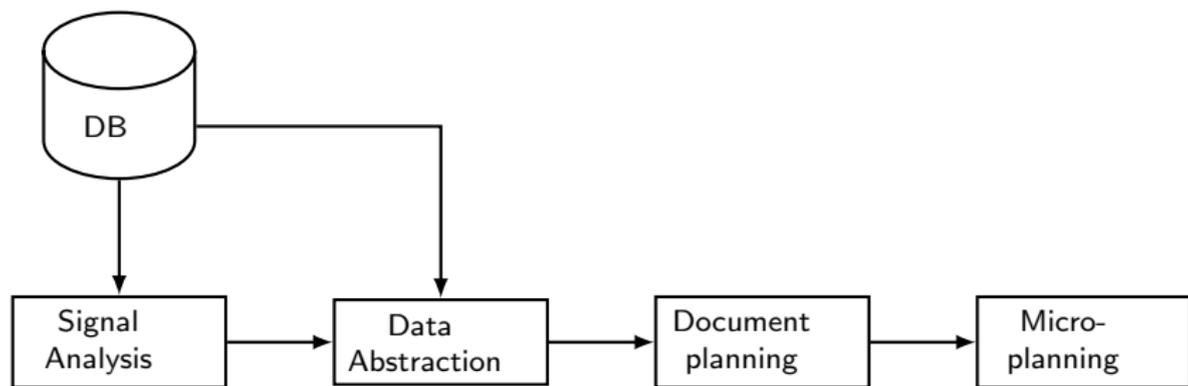- ▶ How to evaluate ?

Architecture

What are the several steps to go from data to text ?

Signal processing $+$ data abstraction $+$ classical NLG pipeline

Who to make sure that a piece of data has the same meaning all along the process ?

Architecture

What are the several steps to go from data to text ?

Signal processing + data abstraction + classical NLG pipeline

Who to make sure that a piece of data has the same meaning all along the process ?

Architecture

What are the several steps to go from data to text ?

Signal processing + data abstraction + classical NLG pipeline

Who to make sure that a piece of data has the same meaning all along the process ?
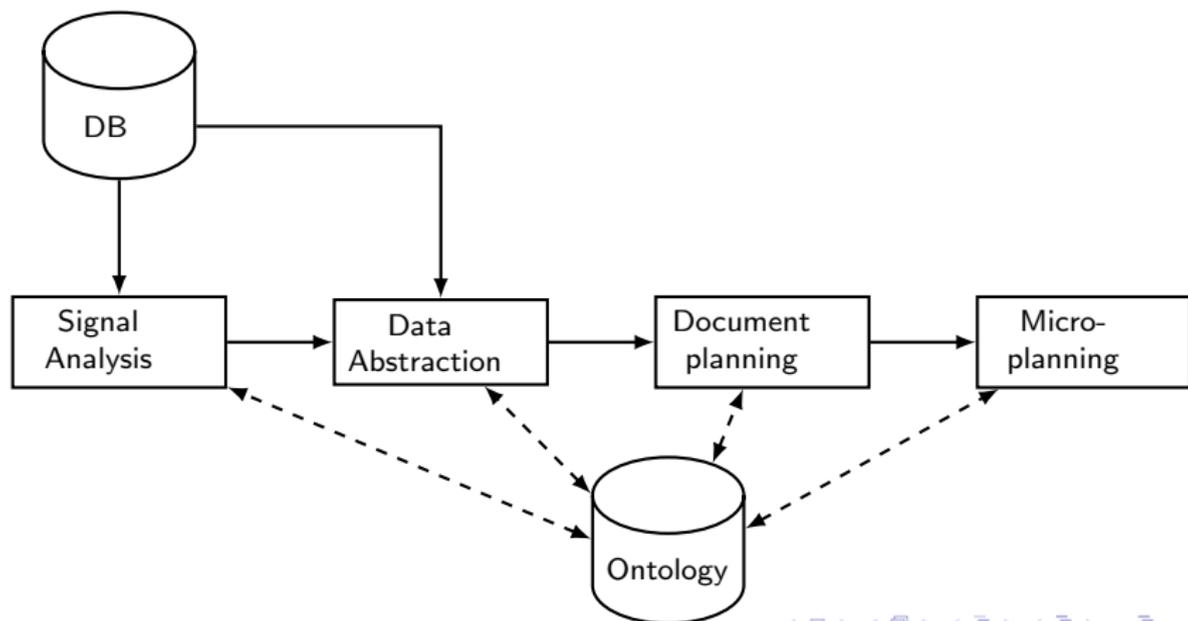
Ontology role in BabyTalk

Off-line
Help to structure our reasoning and to agree on concepts

On-line
Share the concepts (and instances) between the different modules during processing.
Contain knowledge used by the different modules (e.g. valid range of a channel, name of a channel)

Kinds of ontology
BT-45 : Protégé-Frame : Java, knowledge base to store instances, JessTab : about 550 classes
BT-Nurse : OWL : Java, knowledge base to store instances, home-made API to access and reason with the ontology : about 1000 classes

How to select and structure information ?

data abstraction can generate thousand of events $\rightarrow$ a summary can report about 20 events.

In BT-45

- ▶ each piece of data has an *importance* coefficient
- ▶ importance is computed through expert knowledge and heuristics
- ▶ selection of the head events and their context events
- ▶ stop criterion given by summary size

medically important $\neq$ important to report.

How to generate correct medical English ?

- ▶ Lexicalisation : rule-based mapping of EVENTs to predicates based on their ontological class
- ▶ Verbnet, BT-NURSE also used the Specialist Lexicon (UMLS)
- ▶ temporal coherence maintained by the discourse manager
- ▶ realisation performed using SimpleNLG
  [Gatt and Reiter, 2009]
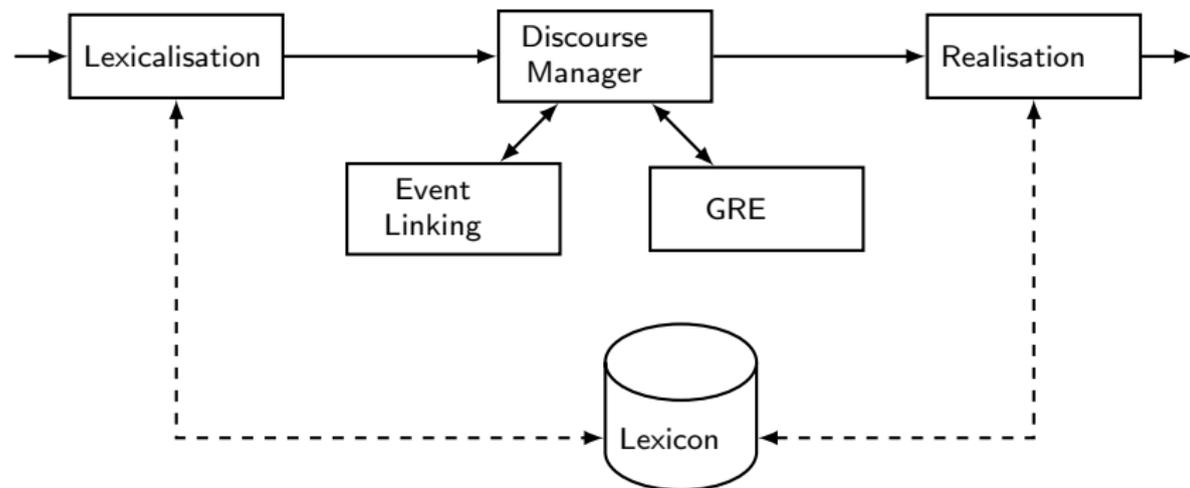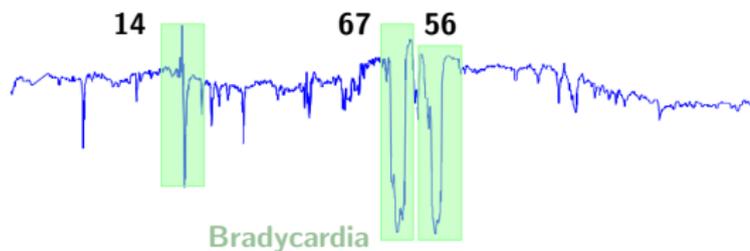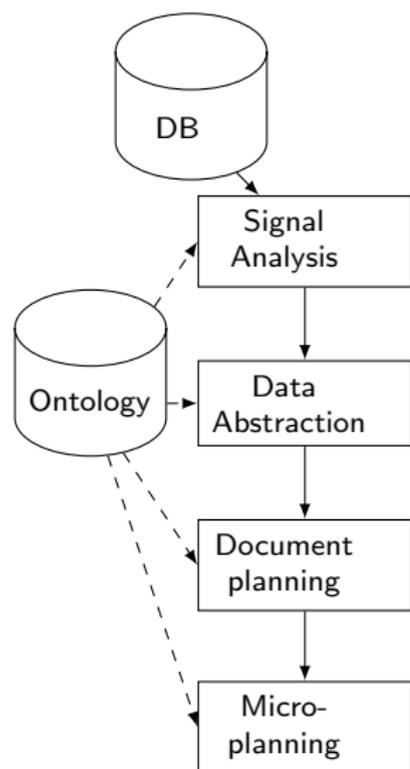
BT principle in one slide!
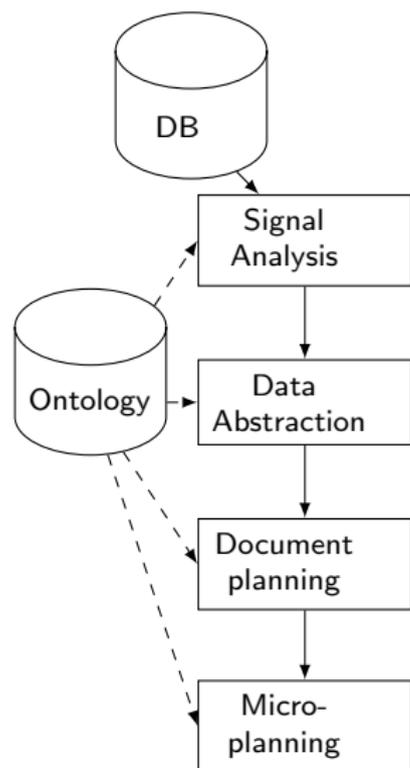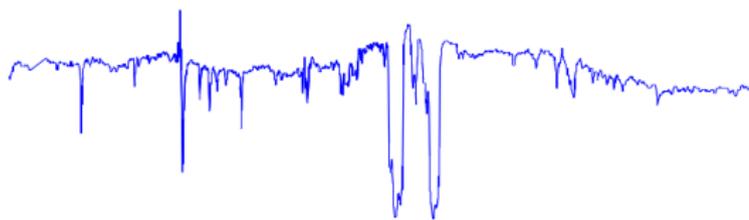
BT principle in one slide !

BT principle in one slide !
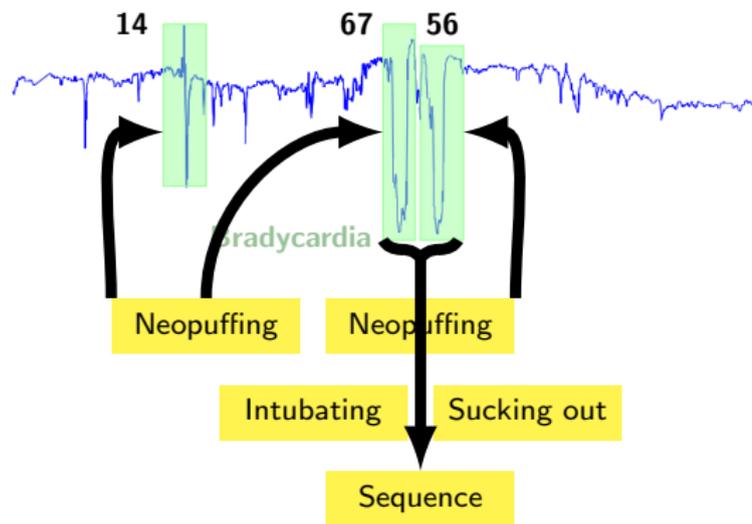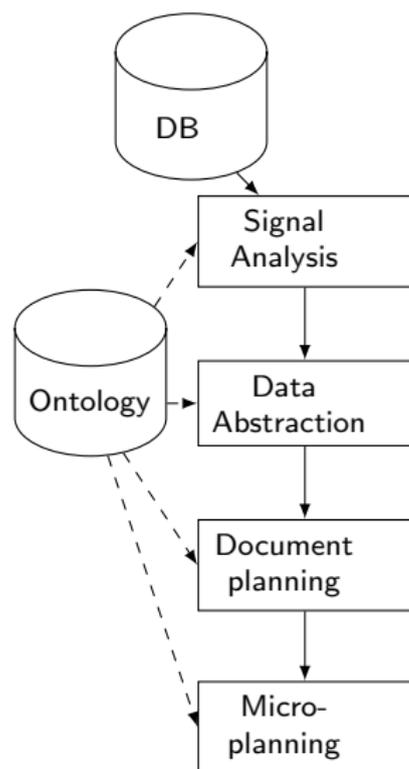
BT principle in one slide !
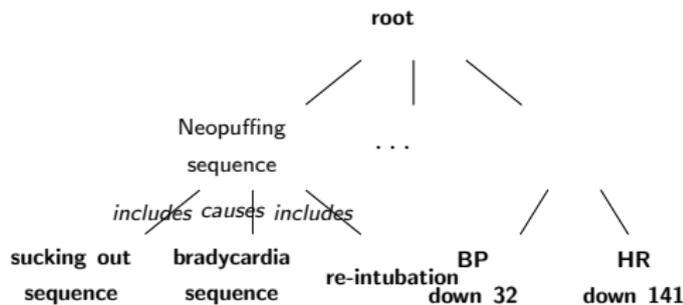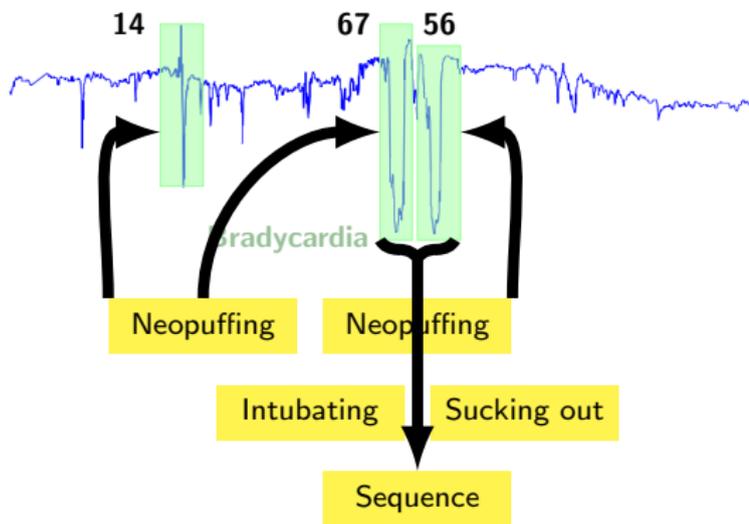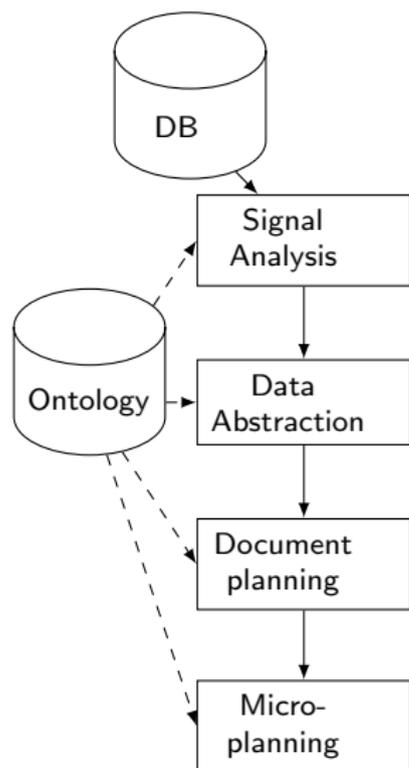
BT principle in one slide !
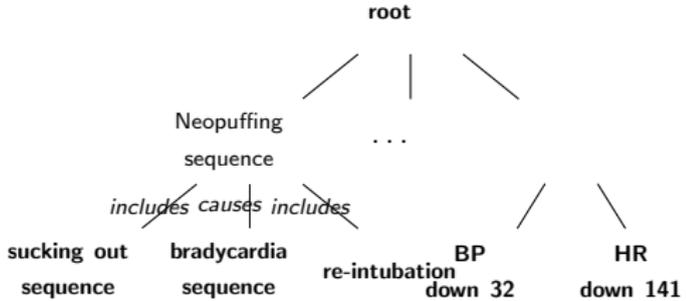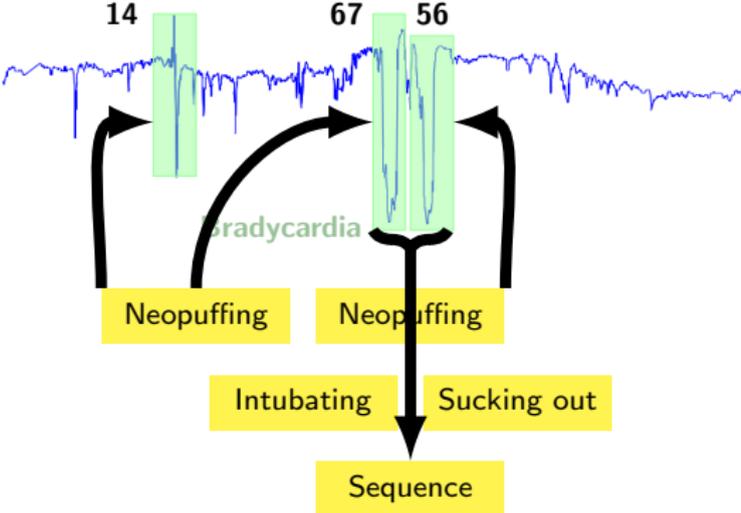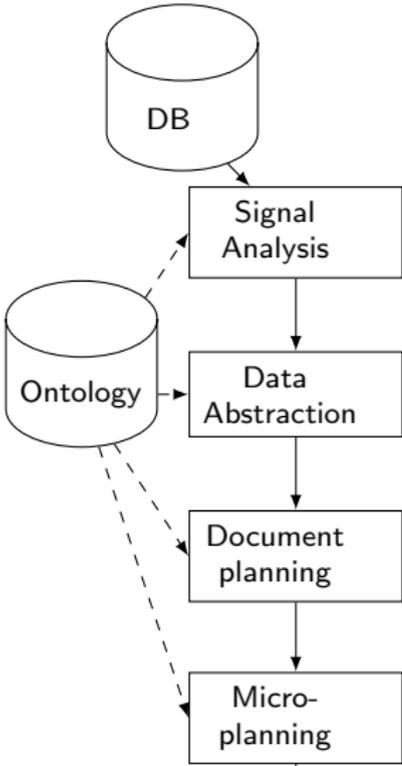
"By 11 :00 the baby received Neopuff ventilation a number of times causing 2 successive bradycardias. After 2 attempts she was re-intubated successfully. The baby was sucked out twice."

How to evaluate ?

Extrinsic methods (looking at a system's utility) :

- ▶ heavily dependent on domain and purpose
- ▶ don't tell us which aspects of the generation process enhance effectiveness (end-to-end evaluation).
- ▶ Task based experiment extremely costly

Intrinsic methods (looking at output quality in its own right) :

- ▶ human judgements tell us what users like, but not necessarily "what's good for them". (Law et al 2005)
- ▶ Automatic measure quick but poorly informative

The relationship between these has been questioned in many areas of NLP (Dorr et al 2005, Belz and Reiter 2010, Callison-Burch et al 2006)

Domain knowledge-based metrics shed more light on the relationship between deep semantic properties of a text and task performance [Gatt and Portet 2009], but stay at a research stage.

BT-45 controlled Task Oriented Evaluation

- ▶ Participants : 35 neonatal ICU nurses and doctors (juniors and seniors).
- ▶ 24 scenarios of about 45 minutes of data.
- ▶ Asked what action(s) they would take at the end of the period.
- ▶ 8 main actions out of 18 possible actions
- ▶ For each scenario, actions could be appropriate, neutral, or inappropriate
- ▶ 3 conditions of presentation :
  - G : Graphic
  - H : Human expert texts
  - C : Computer (BT-45) generated texts

Experiment took place off-ward.

BT-45 controlled Task Oriented Evaluation

- ▶ Participants : 35 neonatal ICU nurses and doctors (juniors and seniors).
- ▶ 24 scenarios of about 45 minutes of data.
- ▶ Asked what action(s) they would take at the end of the period.
- ▶ 8 main actions out of 18 possible actions
- ▶ For each scenario, actions could be appropriate, neutral, or inappropriate
- ▶ 3 conditions of presentation :
  - G : Graphic
  - H : Human expert texts
  - C : Computer (BT-45) generated texts

Experiment took place off-ward.

Exit Contexts Catalogues Displays Studies Tools Help

Catalogue | Display | Configure plots | Configure displays

Clock ..   Yr: 0101 Dy: 24 17:25:00   DataPeriod   → Maximum

TextPlot: BG Text

TemporalPlot: GVT-Obs   TextPlot: H Text   TextPlot: HSEG Text   TextPlot: C Text

BACKGROUND

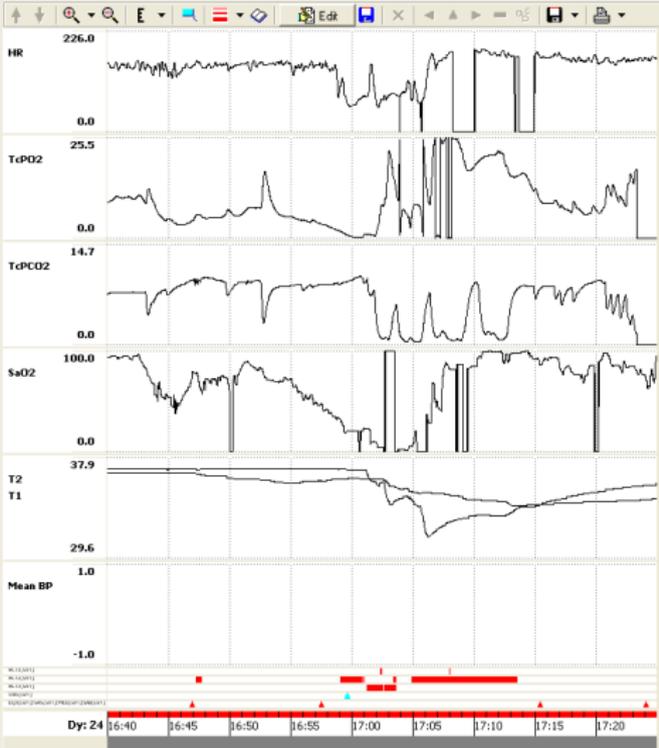Born at 24 weeks and 3 days gestation, birth weight 770 grams, now 22 days old.

Ventilated initially, extubated on day 3 to CPAP. Three days ago developed sepsis and was re-ventilated. He is receiving a combination of IV fluids and enteral feeds. Current treatment includes antibiotics and phosphate supplements, and a top-up transfusion is in progress.

He is in an incubator set at 35.8ºC. Ventilator settings are CMV, rate 15, pressures of 16 / 4, iT is 0.3 seconds, in 70% oxygen.

A capillary blood gas was taken at 16:15:

pH	7.332
pCO$_2$	6.9
pO$_2$	5.04
HCO$_3$	26.2
BE	-0.7

HR   226.0 — 0.0
TcPO2   25.5 — 0.0
TcPCO2 14.7 — 0.0
SaO2   100.0 — 0.0
T2 T1   37.9 — 29.6
Mean BP 1.0 — -1.0

Dy: 24 16:40 16:45 16:50 16:55 17:00 17:05 17:10 17:15 17:20

What should be done?   ACCEPT   78%

- [ ] No action
- [ ] Adjust ventilation / FiO2
- [ ] All cares
- [ ] Blood transfusion
- [ ] Calm / comfort the baby
- [ ] Check / adjust CPAP
- [ ] Commence CPR
- [ ] Extubate
- [ ] Give surfactant
- [ ] Insert chest drain
- [ ] Intubate
- [ ] Manage temperature
- [ ] Monitoring equipment
- [ ] Septic screen
- [ ] Suction
- [ ] Support blood pressure
- [ ] Take blood gas
- [ ] X-ray

Time Series Workbench (9.3) User: ER Catalogue: GraphVsText

Exit Contexts Catalogues Displays Studies Studies Tools Help

Catalogue Display Configure plots Configure displays

Clock...   Yr: 0101 Dy: 24 17:25:00   DataPeriod   → Maximum

TextPlot: BG Text

TemporalPlot: GVT-Obs   TextPlot: H Text   TextPlot: HSEG Text   TextPlot: C Text

BACKGROUND

Born at 24 weeks and 3 days gestation, birth weight 770 grams, now 22 days old.

Ventilated initially, extubated on day 3 to CPAP. Three days ago developed sepsis and was re-ventilated. He is receiving a combination of IV fluids and enteral feeds. Current treatment includes antibiotics and phosphate supplements, and a top-up transfusion is in progress.

He is in an incubator set at 35.8ºC. Ventilator settings are CMV, rate 15, pressures of 16 / 4, iT is 0.3 seconds, in 70% oxygen.

A capillary blood gas was taken at 16:15:

pH	7.332
pCO$_2$	6.9
pO$_2$	5.04
HCO$_3$	26.2
BE	-0.7

You saw the baby between 16:40 and 17:25. Heart Rate (HR) = 155. Core Temperature (T1) = 36.9. Peripheral Temperature (T2) = 36.6. Transcutaneous Oxygen (TcPO2) = 9.0. Transcutaneous CO2 (TcPCO2) = 7.4. Oxygen Saturation (SaO2) = 94.

Over the next 24 minutes there were a number of successive desaturations down to 0. Fraction Inspired Oxygen (FIO2) was raised to 100%. There were 3 successive bradycardias down to 69. Neopuff ventilation was given to the baby a number of times. The baby was re-intubated successfully. The baby was resuscitated. The baby had bruised skin.

Blood gas results received at 16:45 showed that PH = 7.3, PO2 = 5, PCO2 = 6.9 and BE = -0.7.

At 17:15 FIO2 was lowered to 33%. TcPO2 had rapidly decreased to 8.8. Previously T1 had rapidly increased to 35.0.


What should be done?   ACCEPT

5%

☐ No action	☐ Calm / comfort the baby	☐ Give surfactant	☐ Monitoring equipment	☐ Take blood gas
☐ Adjust ventilation / FiO2	☐ Check / adjust CPAP	☐ Insert chest drain	☐ Septic screen	☐ X-ray
☐ All cares	☐ Commence CPR	☐ Intubate	☐ Suction	
☐ Blood transfusion	☐ Extubate	☐ Manage temperature	☐ Support blood pressure	

Results of the BT-45 Evaluation

$$score = \frac{|\text{chosen app actions}|}{|\text{app actions}|} - \frac{|\text{chosen inapp actions}|}{|\text{inapp actions}|} \qquad (1)$$

$G = .33^{.14}$
$H = .39^{.11}$
$C = .34^{.14}$
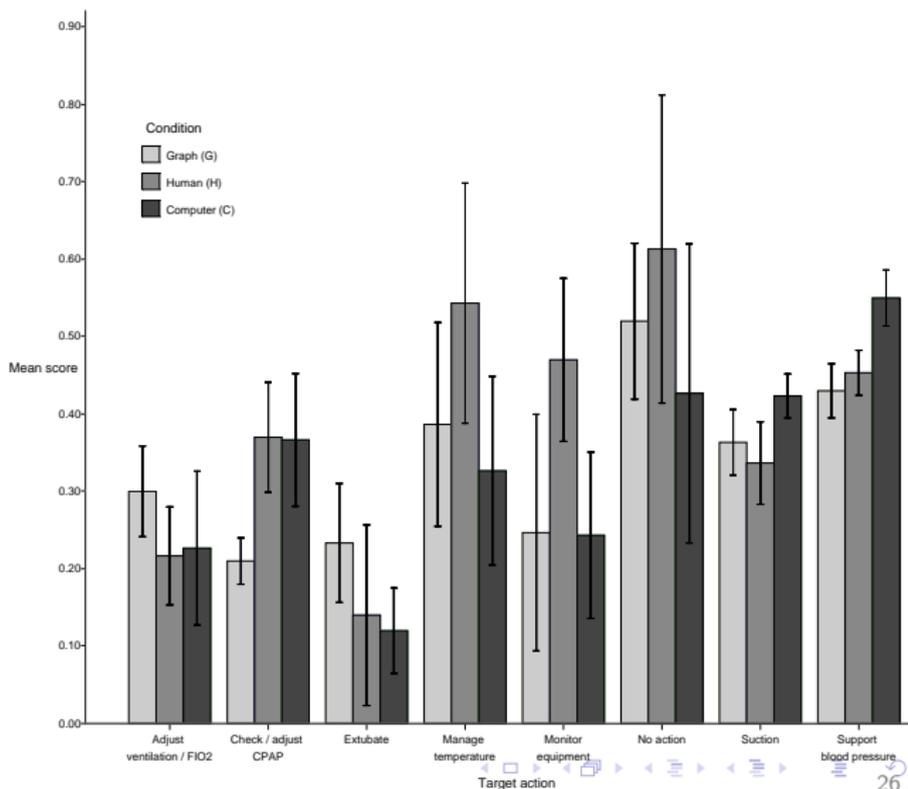
Conclusion

Generating good quality and effective summaries from heterogeneous multimodal inaccurate temporal data is possible !
Strong link between NLG and multimedia generation [André, 2000]

Issues

- Lack of Contextual Importance (BT-45)
- Temporal Issues : Difficulty of handling short and long term events, Discontinuity
- Narrative structure (making stories)
- Still difficult to evaluate

Thank You

The babytalk project :
www.csd.abdn.ac.uk/research/babytalk/

📄 A. Gatt and F. Portet. (2010)

Textual properties and task-based evaluation : Investigating the role of surface properties, structure and content. In *Proc. INLG 2010*.

📄 F. Portet and A. Gatt. (2010)

Towards a Possibility-Theoretic Approach to Uncertainty in Medical Data Interpretation for Text Generation. In *Proc. Knowledge Representation for Health-Care : Data, Processes and Guidelines*.

📄 F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. (2009)

Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7–8) :789–816.

References

L. Danlos and L. Roussarie. (2000)
La génération automatique de textes. In *J.M. Pierrel (éd), Ingénierie de la langue, Hermès*.

D. Hüske-Kraus. (2003)
Suregen-2 : A shell system for the generation of clinical documents. In *EACL-03*.

E. Reiter, R Robertson and L Osman. (2003)
Lessons from a Failure : Generating Tailored Smoking Cessation Letters. In *Artificial Intelligence*.

C. Callaway and J. Lester. (2002)
Narrative prose generation. In *Artificial Intelligence*.

A. Gatt and E. Reiter. (2009)
SimpleNLG : A realisation engine for practical applications. In *ENLG-2009*.

E. André. (2000)
The generation of multimedia presentations In *Handbook of Natural Language Processing : Techniques and applications for the processing of language as text*.