

Natural Language Generation and Data-To-Text

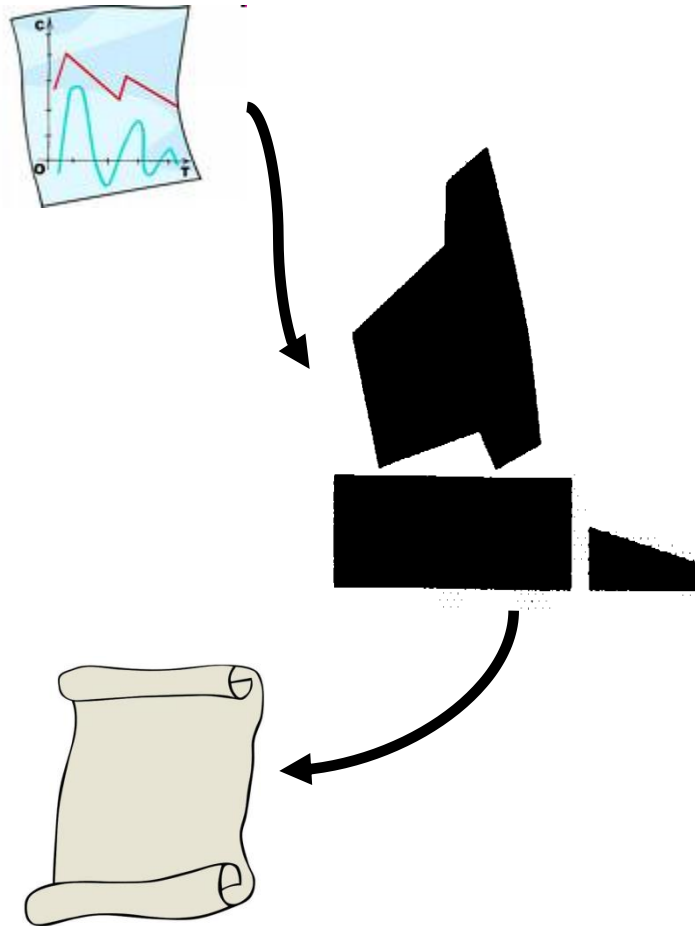
Albert Gatt

Institute of Linguistics, University of Malta

Tilburg center for Cognition and Communication (TiCC)

Department of Computing Science, University of Aberdeen

A typical evaluation scenario



How good is my generator?

- Do people **like** the output?
- Does it **compare well to what people do**?
- Does it **help people achieve a task**?

Are these questions completely different, or are they related in some way?

Evaluation in NLG

Evaluation strategies

- Evaluate performance of single modules
 - e.g. realisation
 - e.g. Generation of Referring Expressions
- Evaluate entire system (“end to end evaluation”).

Evaluation methods (Sparck Jones & Galliers `95)

- Extrinsic (looking at a system’s utility)
 - through experiments with human users in relevant settings
 - “Does the system/module achieve what it is meant to achieve?”
- Intrinsic (looking at output quality in its own right)
 - against corpora
 - “Is the output like that produced by people in a similar situation?”
 - by eliciting human judgements
 - “Do people like the output?”

NLG Evaluation: extrinsic/task-oriented

Method

- Evaluation with target users
 - STOP: how effective are generated letters in motivating people to stop smoking? (Reiter `05)
 - PEACH: are people's museum visits enhanced by automatic generation of information on a PDA? (Stock et al `07)
 - BT-45: do generated summaries help doctors and nurses take clinical decisions? (Portet et al `09)
- Widely accepted as the most conclusive sort of NLG evaluation.
- Very expensive!
 - STOP evaluation cost ca. £75,000
 - BT-45 evaluation cost ca. £20,000 (Reiter/Belz `09)
 - BT-NURSE evaluation cost even more!

NLG evaluation: intrinsic/human

Method

- Show readers a generated text and ask for judgements
 - JAPE: do children recognise generated jokes as such? Are they funny? (Binsted et al `97)
 - SumTime: how do readers judge generated vs expert-written forecasts? (Reiter et al `05)
 - COMIC: which strategy for combining speech and gesture is preferred by human judges? (Foster/Oberlander `07)
 - STORYBOOK: which components of a story generator contribute to making readers' judgements more positive? (Callaway/Lester `02)
- One of the most widespread methodologies in NLP.
- Probably more appropriate for some domains than extrinsic evaluation.

NLG evaluation: intrinsic/automatic

Method

- Compare output to a corpus of human-authored reference texts using some evaluation metric.
 - Not very typical in end-to-end system evaluation (but see Belz and Kow `09; Reiter and Belz `09)
 - To date, mainly used to evaluate coverage and correctness of realisers (Langkilde-Geary `02; Callaway `03)

Precedents

Intrinsic evaluation in Machine Translation

- Heavily dominated by BLEU (Papineni et al `02)
 - Essentially, intrinsic, corpus-based method;
 - Often used in NLG evaluations of realisers.
- Some doubts have been cast on whether BLEU properly reflects translation quality (Callison-Burch et al `06).
 - E.g. to what extent does it correlate with human judgements?
- Question: which human judgements?
 - It's one thing to ask monolinguals to judge a translation, quite another to ask bilinguals.

Precedents

Intrinsic evaluation of automatic summarisation

- Methods range from judgements (pyramid method) to automatic intrinsic metrics (ROUGE)
- Correlations have been reported between ROUGE and human judgements.
- Very low correlations between ROUGE and extrinsic methods (based on relevance) (Dorr et al `05).

Precedents in NLG

Role of corpora

- Recent evaluations show that judgements of generated texts don't correlate well with corpus-based measures (Reiter/Belz '09).
- Many have questioned whether corpus texts should be treated as NLG “gold standards” (e.g. Reiter/Sripada '02).

Role of judgements

- Experimental results suggest that people's preferences don't reflect their task performance:
 - Law et al '05: compared medical decision making with visualisations vs human-written summaries of patient data.
 - Doctors & nurses preferred visualisations.
 - But decision-making was much better with text!

Why this matters

- Ultimately, we're interested in the relationship between these methods.
 - Do automatic results against corpora tell us anything about the quality of our output?
 - Do people's judgements converge with corpus-based results?
 - Can we predict effectiveness of a text (extrinsic) from corpus-based metrics or judgements?

Part 1

EXTRINSIC EVALUATION OF THE BABYTALK BT-45 PROTOTYPE SYSTEMS

BabyTalk Evaluation

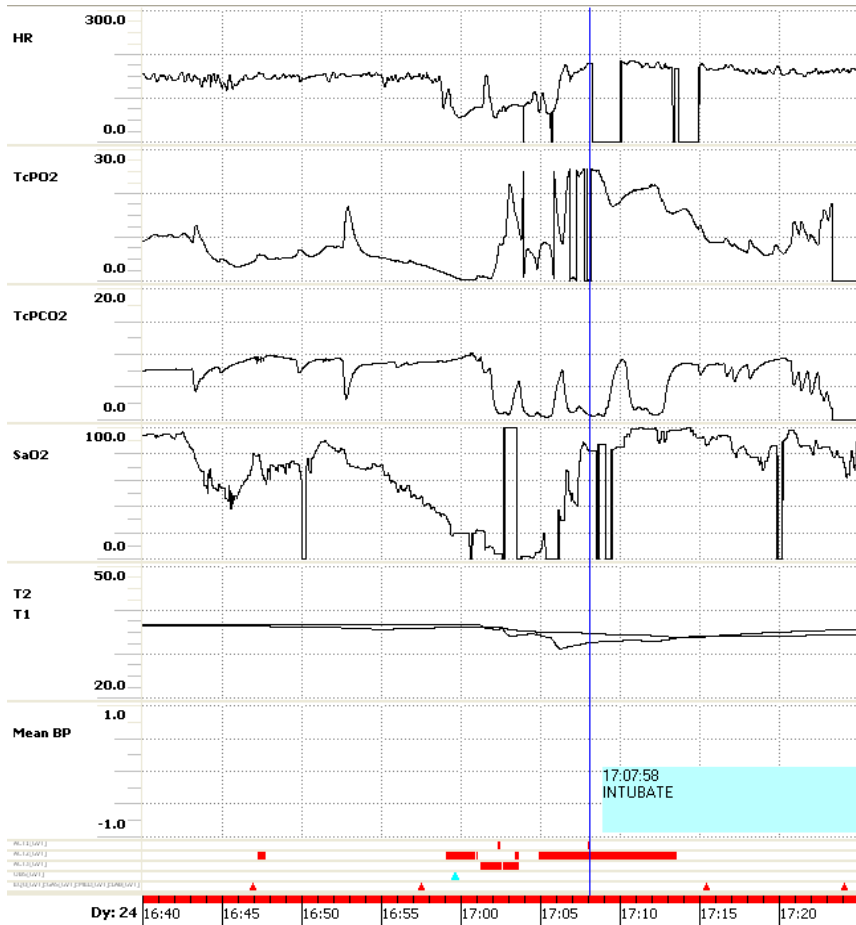
- In BabyTalk, there was a fairly clear target user population.
- Aim was also fairly clear: help medics (doctors, nurses) take clinical decisions.
 - BT-Family had a different aim.
- Here, we'll look at:
 - Evaluation of BT-45: prototype system which summarises 45 minutes of data.
 - Evaluation of BT-Nurse: system that summarises 12 hours of data (next part)

BT-45 Summary

Input data



Output text



[...]

Over the next 24 minutes there were a number of successive desaturations down to 0. Fraction of Inspired Oxygen (FIO2) was raised to 100%. There were 3 successive bradycardias down to 69.

[...]

Prototype evaluation: BT-45

Method

- Tested decision-making by medical staff given exposure to some information about a patient;
- Experiment conducted off-ward in 2008-9.

Conditions

- **(G)** Data presented as visualisation (the usual way)
- **(H)** Human summary produced by experts
 - consensus summary of the data by a senior neonatal nurse and a consultant neonatologist
- **(C)** Computer-generated summary by BT-45
 - completely automatic

Prototype evaluation: BT-45

Participants

- 35 junior and senior doctors and nurses

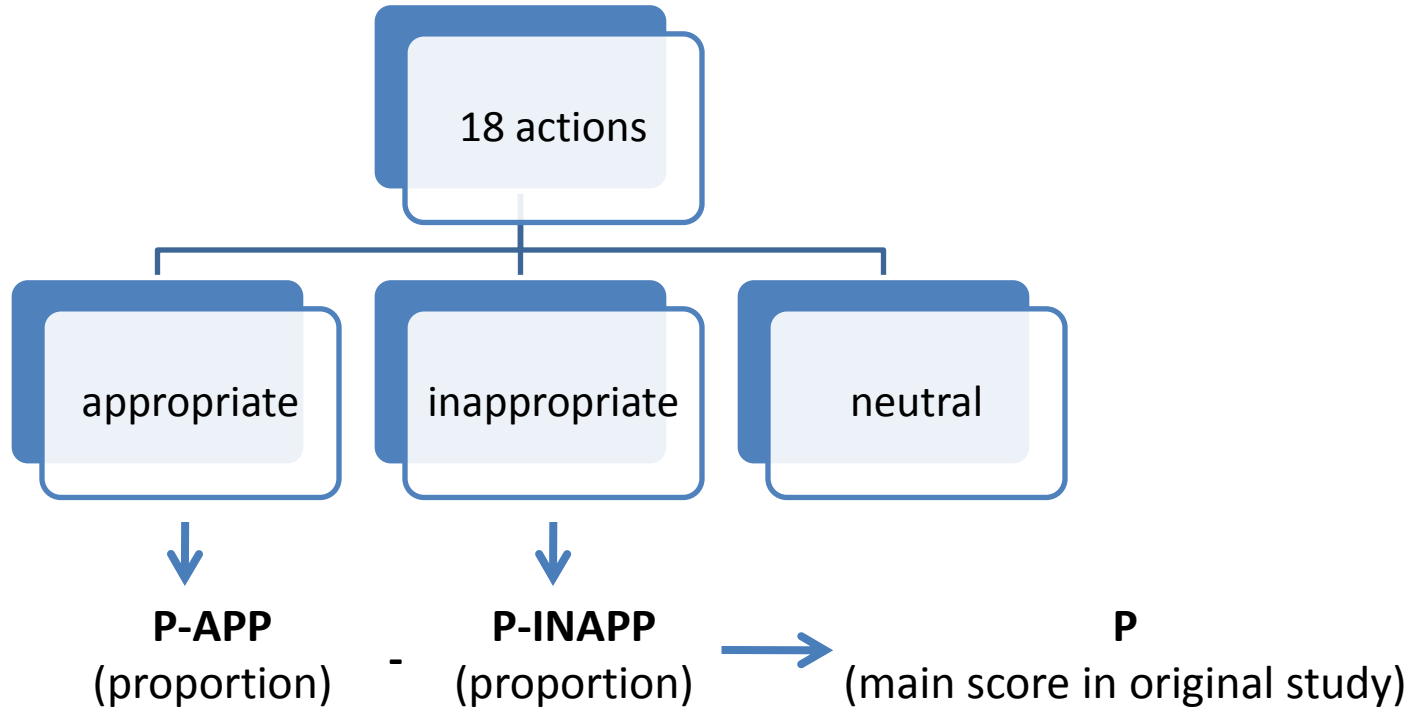
Procedure

- Participants shown 24 scenarios in one of the conditions;
- Asked to select one or more appropriate actions given the summary.
- For each scenario, there were 18 actions to choose from.

Scoring

- Decision-making score reflects the extent to which a participant made appropriate selections (range: [0,1]).

Performance metrics



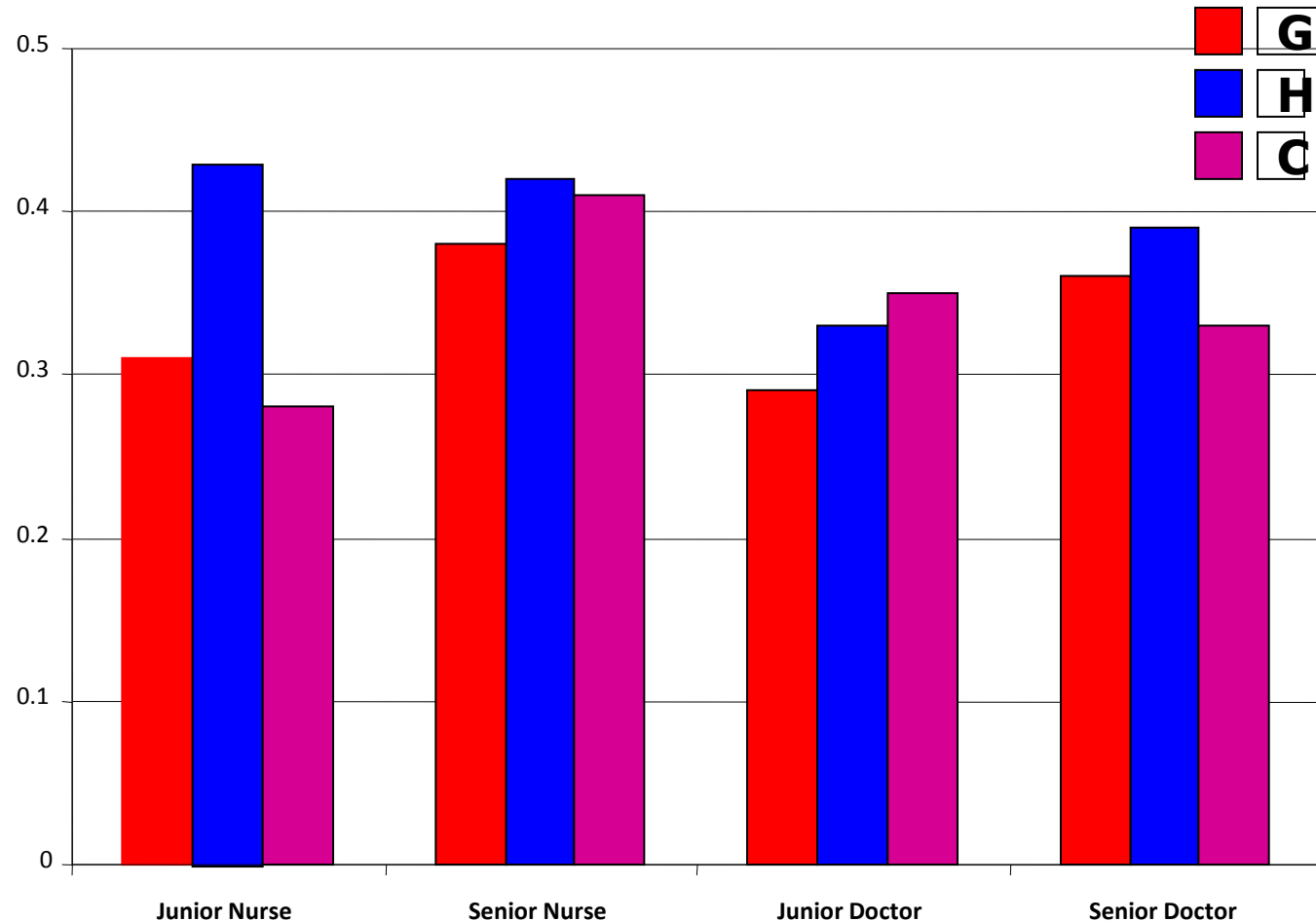
Prototype Evaluation: Results

G = 0.33 ^{0.14}
H = 0.39 ^{0.11}
C = 0.34 ^{0.14}

**Biggest difference
for Junior Nurses**

H best overall

G no better than C



Prototype Evaluation: Summary

- Generated summaries are at least as good as graphical presentations.
- Automatic summarisation shown to be feasible in the neonatal context.
- Text may be more effective for some user groups (Junior Nurse).
- But what makes human texts better?

BT-45 vs Human

A qualitative comparison

- BT-45 texts suffered from a lack of continuity.
 - **BT-45:** *TcPO2 suddenly decreased to 8.1. [...] TcPO2 suddenly decreased to 9.3.*
- BT-45 focus is on medically important events. Humans take context into account.
 - **Human:** *The pO2 and pCO2 both have spikes in the traces [...] There are several episodes of artefact*
 - **BT-45:** never mentions noise or artefact.
- BT-45 doesn't give long-term overviews. Humans do this all the time.
 - **Human:** *He is warm centrally.*
 - **BT-45:** *Core Temperature (T1) = 36.4*
- Humans handle time much better!

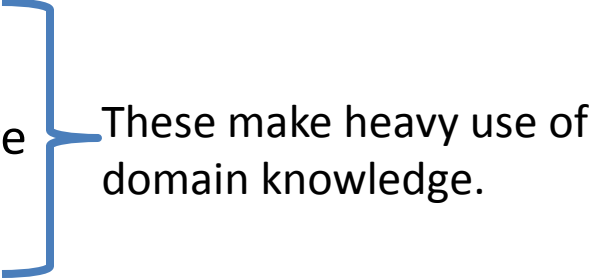
Question

Is there a relationship between the content of the generated texts and the level of decision making by doctors and nurses?

- We tried to answer this in an evaluation that tried to predict decision-making performance from the textual content.
- This relied very heavily on the domain knowledge in the system.

Can we quantify this somehow?

Our method

- Compare human and BT-45 texts on a number of metrics based on
 - surface structure
 - content
 - discourse/temporal structure
 - Relevance
- 
- These make heavy use of domain knowledge.

In practice

- We annotate the human and BT-45 texts, using our ontology.
- We compute scores, most of which rely heavily on that ontology.
- We see if there is a relationship between the text scores and the decision-making scores in the experiment.

Textual annotation I: Content

1. Using **the ontology**, human and BT-45 texts were annotated with the **events** they express.
2. Annotation also reflects the **functional concept (roughly, body system)** that an event in the text is associated to.

*there are **brief desaturations to the mid 80s**;*

TYPE = **DESATURATION**
ASSOCIATED-TO = **RESPIRATION**

*recovery to baseline is **spontaneous***

TYPE = **TREND**
SOURCE = **Saturation O2**
ASSOCIATED-TO = **RESPIRATION**

Textual annotation II: Structure

- **At 14:15** hours a heel prick is done.

TREL: *at*
ARG0: HEEL-PRICK
ARG1: TIMESTAMP

- The HR increases **at this point** ...

TREL: *starts*
ARG0: TREND(HR)
ARG1: HEEL-PRICK

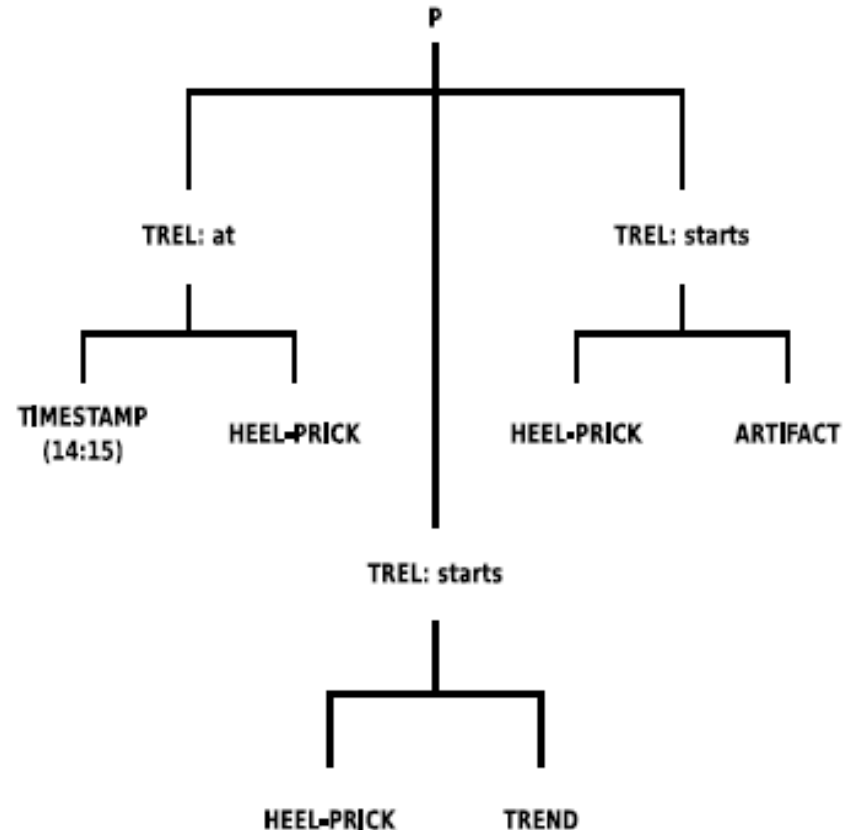
- ... and **for 7 minutes from the start** of this procedure there is a lot of artefact...

TREL: *starts*
ARG0: HEEL-PRICK
ARG1: ARTEFACT

- **Temporal relations** are based on Allen's ('84) typology.
- Also annotated **discourse relations** of CAUSE and CONTRAST (Mann & Thompson '88)

Textual annotation III: deriving trees

- **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 artefact...



Evaluation metrics

Surface properties (n-gram overlap)

- ROUGE-4
- ROUGE-SU (skip bigrams + unigrams)

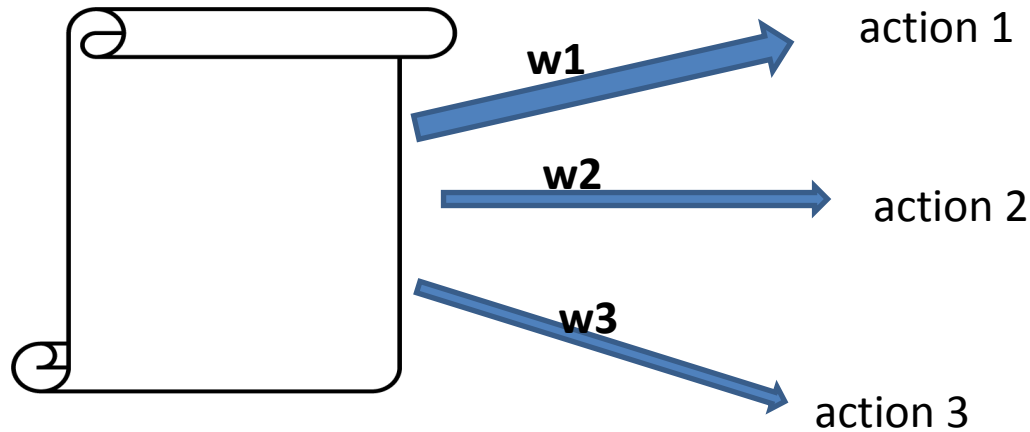
Content and structure

- No. of events mentioned in a text
- No. of temporal/discourse relations mentioned in a text
- Tree Edit Distance between texts

Knowledge-based relevance measure

- REL: extent to which events mentioned in a text are relevant to one or more appropriate action(s).
- IRREL: extent to which events are relevant to one or more inappropriate action(s).

Quantifying relevance – I



Main hypothesis

- Texts can reference actions to different degrees. This may **bias readers towards some actions** more than others.
- This is reflected by a weight assigned to each action, for each text.

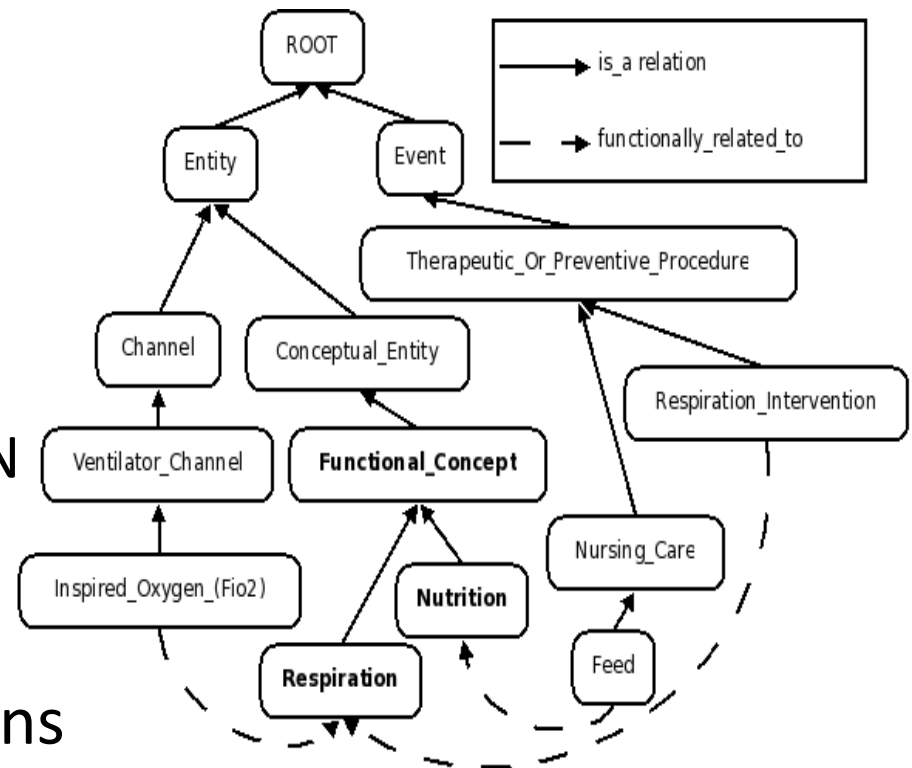
The scores

- REL = sum of weights of appropriate actions referenced in the text
- IRREL = sum of weights of inappropriate actions referenced in the text

Quantifying relevance - II

“There are 3 episodes of desaturation to 70%”

- DESATURATION is functionally linked to RESPIRATION
- So are some of the actions
 - E.g. MANAGE VENTILATION
- The DESAT **references** RESPIRATION-related actions



Quantifying relevance – III

Pruning spurious connections

- With the reasoning so far, we can end up with a lot of actions related to each event.
- But not all of these are warranted, given the status of a patient.
- Solution: use knowledge-based expert rules to prune these connections.

$$\text{INTUBATE-BABY} \leftarrow \neg \text{ON}(\text{BABY}, \text{CMV})$$

- Only assume a connection between an event and the INTUBATE action if the action can indeed be taken.

Quantifying relevance - IV

Estimating prior probabilities

- Not all actions that could have been chosen are equally probable.
 - MANAGE VENTILATION – routine
 - RESUSCITATE PATIENT – drastic, only in highly critical situations
- Actions are weighted by their prior probability, based on a DB of 48K clinical actions (Hunter et al `03)

Event importance

- Events are also weighted by their importance.
- A resuscitation is much more important than a nappy change.

Results – Surface Properties (I)

Correlation with difference in human performance on H and C texts

	ROUGE-4	ROUGE-SU
P	-.19	.04
P-APP	-.2	.01
P-INAPP	-.01	-.1
SP	-.1	.13

- No correlations are significant.

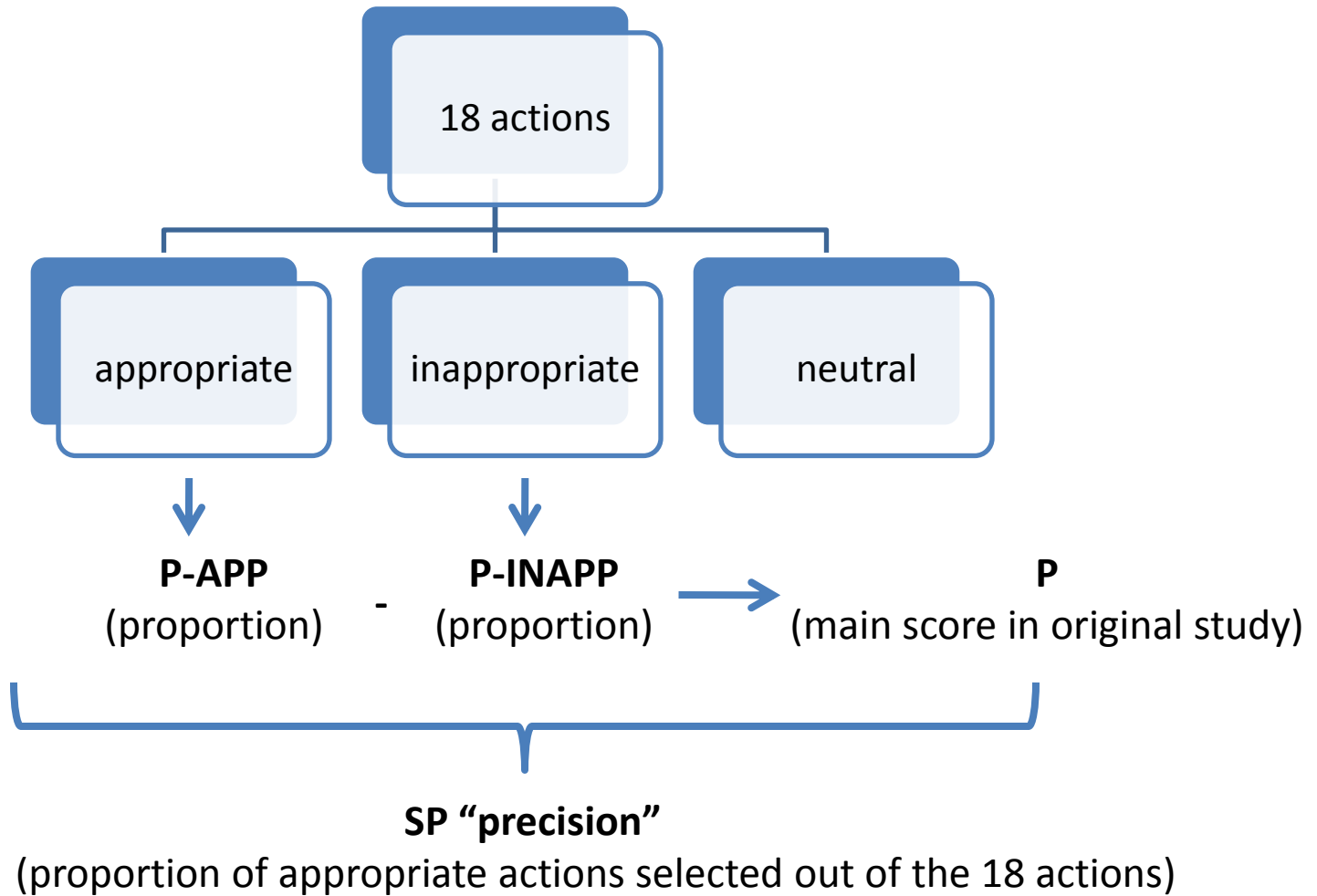
Results – Surface Properties (II)

Correlation with performance on C texts only

	ROUGE-4	ROUGE-SU
P	.33	-.03
P-APP	.38	-.02
P-INAPP	.2	.05
SP	-.03	-.31

- No correlations are significant.

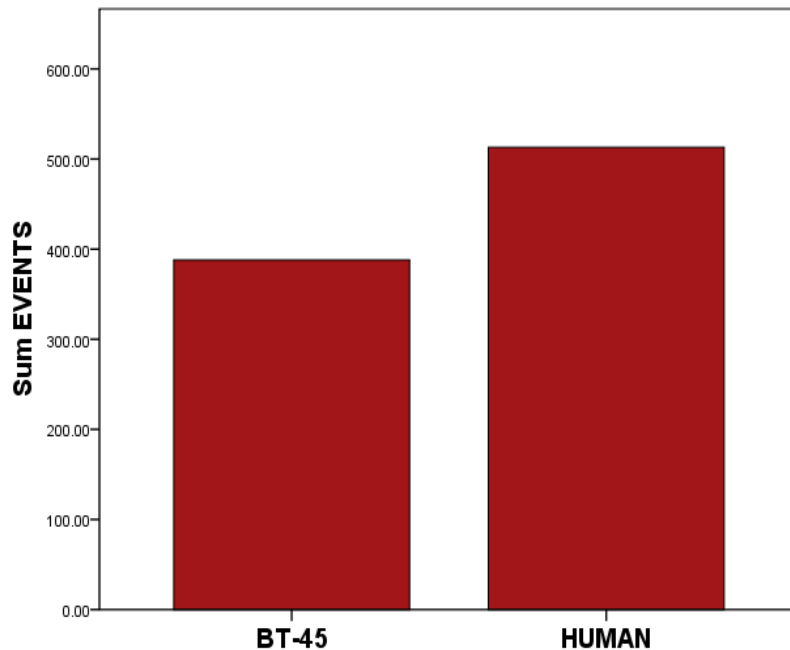
Performance metrics for evaluation



Results – Content and Structure (I)

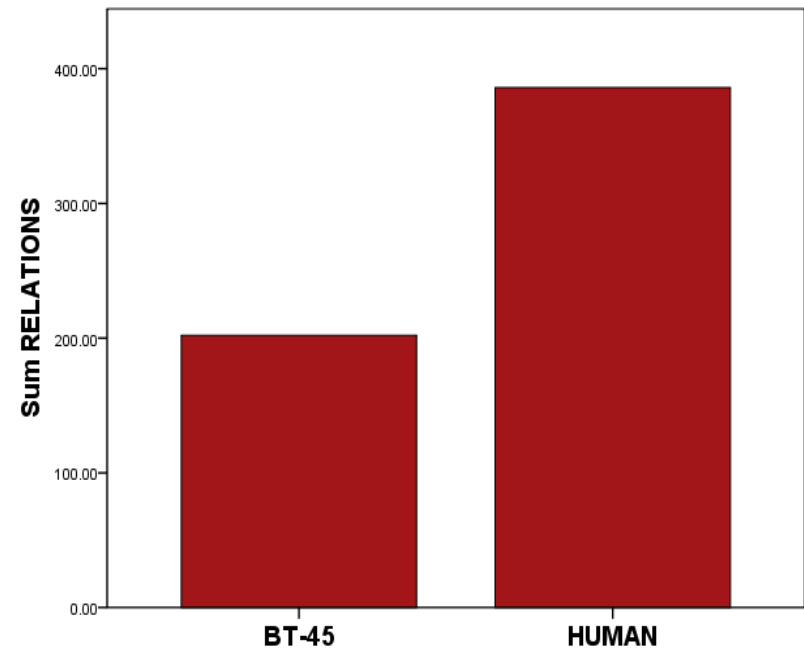
EVENTS

$t = 2.44, p = .05$



RELATIONS

$t = 3.70, p < .05$



Results – Content and Structure (II)

Discourse/temporal relations

- For BT-45 texts:
 - Negative correlation with P-INAP ($r = -.49, p < .05$)
 - Positive correlation with SP precision ($r = .7, p < .001$)
 - Temporal/discourse relations made texts easier to understand in the case of BT-45.

Events

- For BT-45 texts:
 - Negative correlation with P-APP ($r = -.53, p < .05$)
 - Negative correlation with P ($r = -.5, p < .05$)
 - Suggests that BT-45 may have mentioned several “irrelevant” events.

Results – Content & Structure (III)

Correlation with performance differences (H-C)

	EVENTS	RELATIONS	TREE EDIT
P	.43	.34	.36
P-APP	.42	.30	.33
P-INAPP	-.09	-.15	-.14
SP	.02	0	.09

- Performance score is positively correlated to the number of events mentioned.
- Correlations with relations in the predicted direction.
- No significant correlation with Tree Edit
 - Problem with node duplication method?

Results – Relevance

Correlation: Human texts

	REL	IRREL
P	.14	-.25
P-APP	.11	-.22
P-INAPP	-.14	.1
SP	.60	-.56

- All correlations in the predicted direction.
- Positive correlations with our precision score, SP.

Results – Relevance

Correlation: BT-45 texts

	REL	IRREL
P	.33	-.34
P-APP	.24	-.26
P-INAPP	-.49	.43
SP	.7	-.62

- The same overall trends as with H texts.
- Suggests that our knowledge-based relevance metrics may be on the right track.

Summary

- Lack of correlations between “surface” metrics and performance, which echoes similar results by many other authors.
- The content and structure of the text correlates with various performance measures.
 - But we need a better way of comparing underlying text structure directly (graph-based methods?)
- Our experiment also gives us a workable definition of the notion of “relevance”, which is found to correlate with precision on a decision-making task.

Conclusions

- Unlike many previous evaluation studies, we place more emphasis on domain knowledge:
 - Enables us to quantify aspects of content, structure and relevance.
 - These are more systematically related to performance than surface properties.
- From a methodological point of view, our results suggest that evaluating NLG systems must factor in the domain knowledge they incorporate.

Part 3

EVALUATING BT-NURSE

BT-Nurse: Example summary

Respiratory Support

Current Status

Currently, the baby is on CMV in 27 % O₂. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H₂O. Tidal volume is 1.5.

SaO₂ is variable within the acceptable range and there have been some desaturations.

The most recent blood gas was taken at around 07:45. Parameters are acceptable.

Events During the Shift

[...] Another ABG was taken at around 23:00. Blood gas parameters had deteriorated to respiratory acidosis by around 23:00.

[...] The baby was intubated at 00:15 and was put on CMV. [...] He was given morphine and suxamethonium.

[...] Between 00:30 and 03:15, SaO₂ increased from 88 % to 97 %.

Another ABG was taken at around 00:45. pH was 7.18. CO₂ dropped to 7.95 kPa. BE was -4.8 mmol/L.

Evaluating BT-Nurse

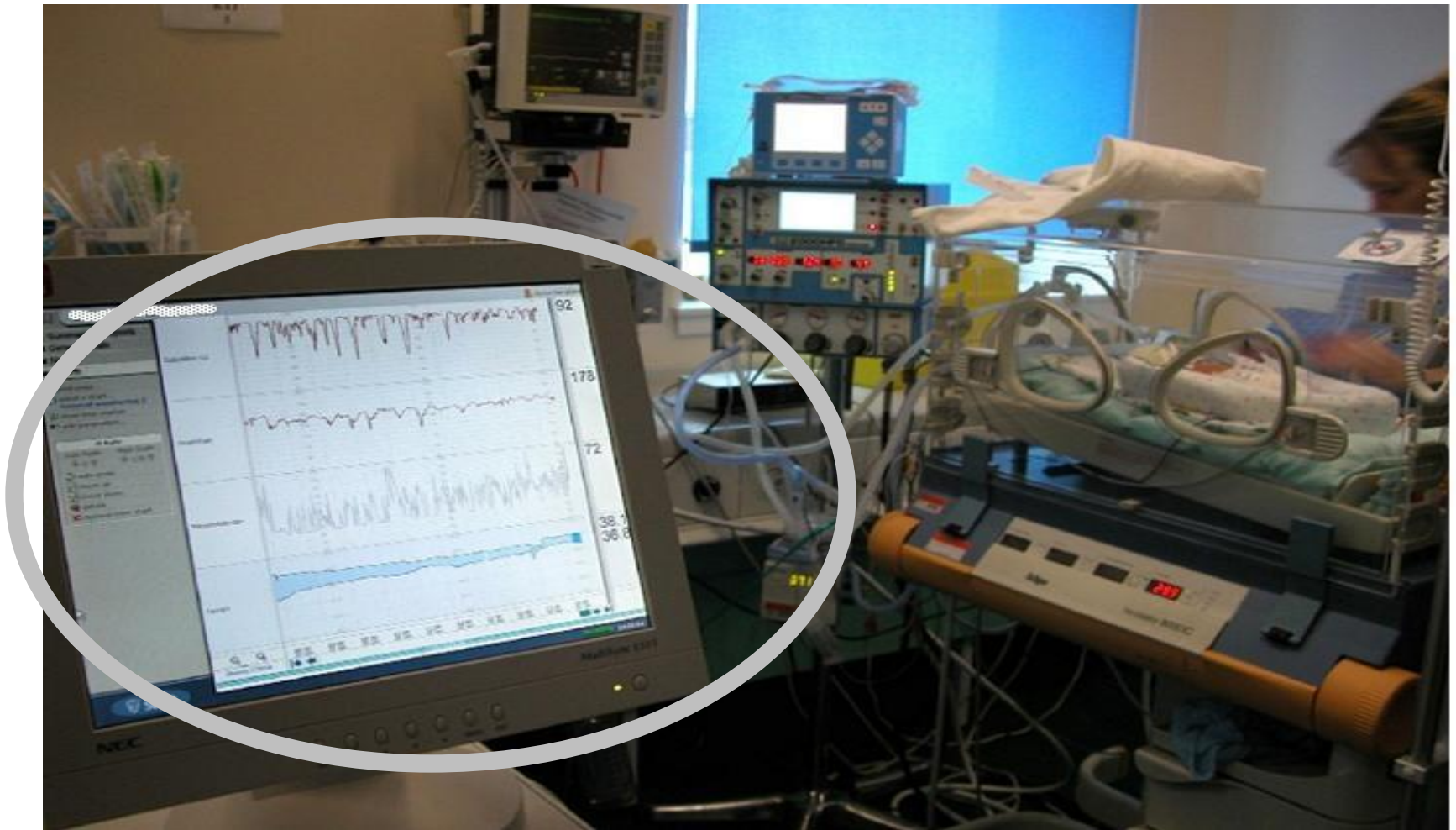
Method

- Deployed on-ward for three months (1 month pilot + 2 months evaluation) at the Edinburgh Royal Infirmary.
 - Integrated with cotside patient management system;
 - Used under supervision (for evaluation purposes);
 - Generated summaries from real-time, unseen patient data.

Details

- 54 different nurses; 31 different patients
- 165 evaluation scenarios
 - 92 with incoming nurses (56%); 73 with outgoing nurses (44%)
- System breakdowns: 0%

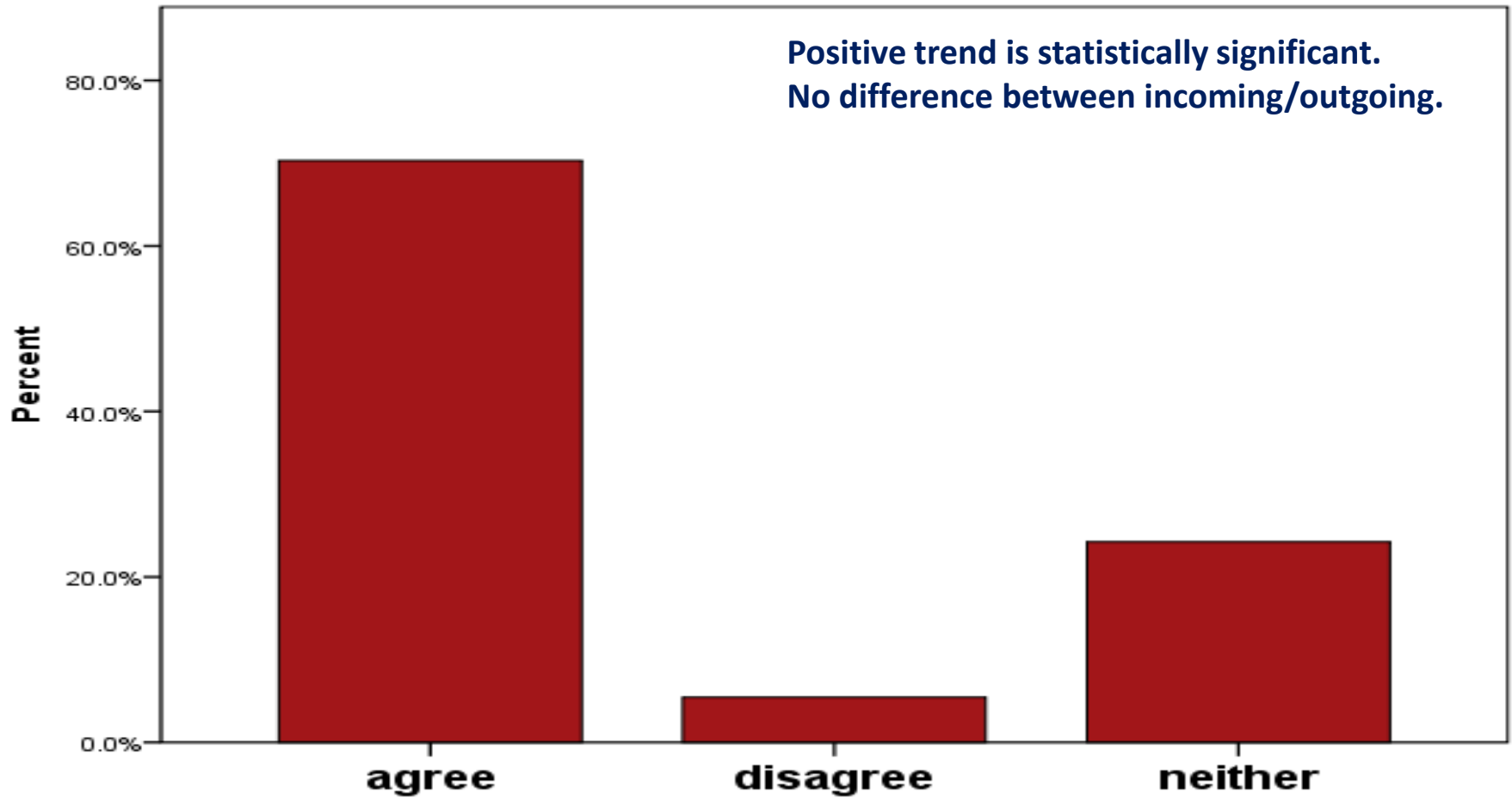
Deployment of BT-Nurse



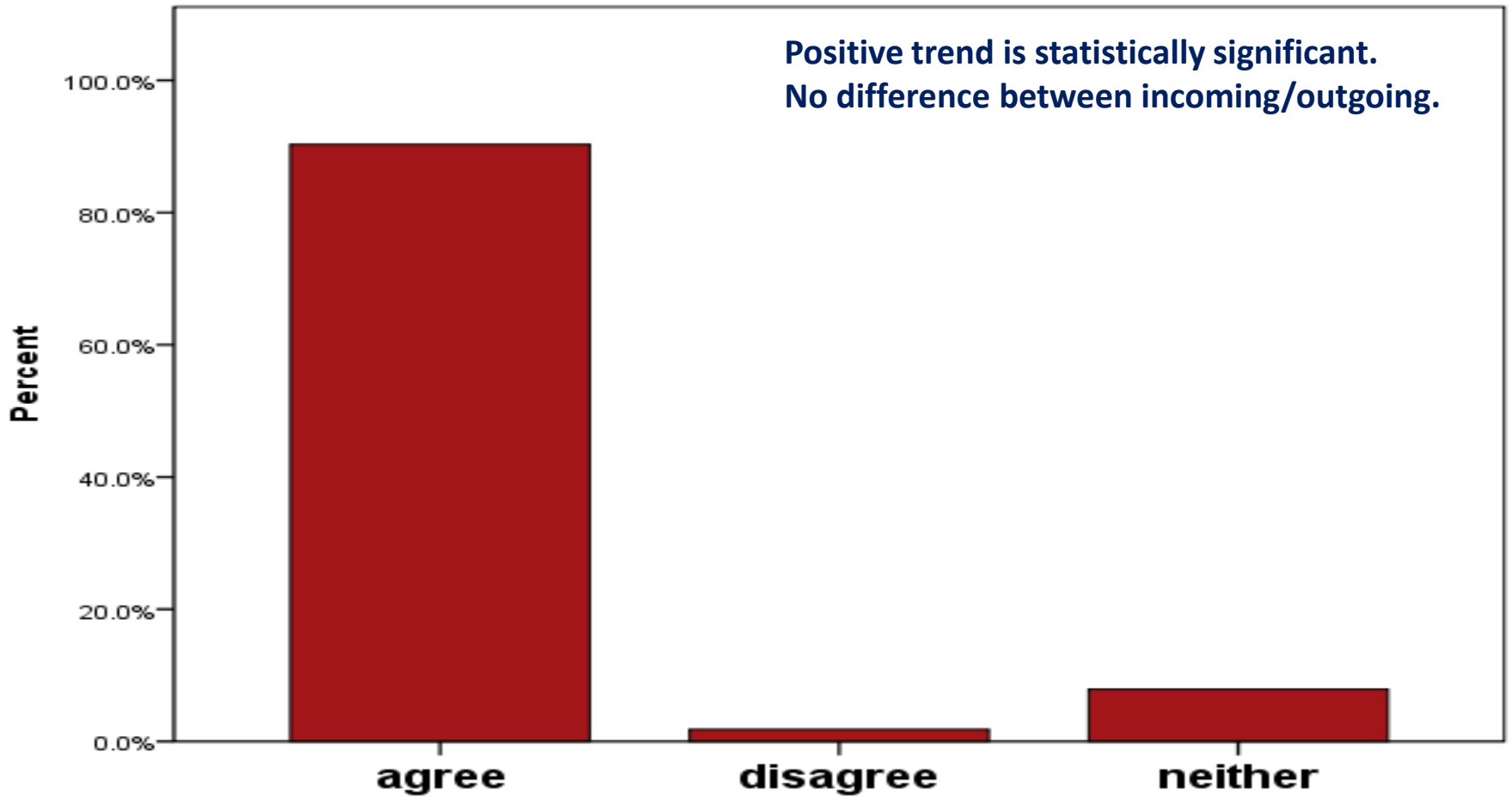
Evaluation measures

- Nurses asked to express dis/agreement with three statements:
 - **Understandability:** "The BT-Nurse summary was easy to understand"
 - **Accuracy:** "The BT-Nurse summary is accurate"
 - **Utility:** "The BT-Nurse summary would help me..."
 - "...write a shift summary" [outgoing]
 - "...plan a shift" [incoming]
- Nurses also asked to comment freely on the text.

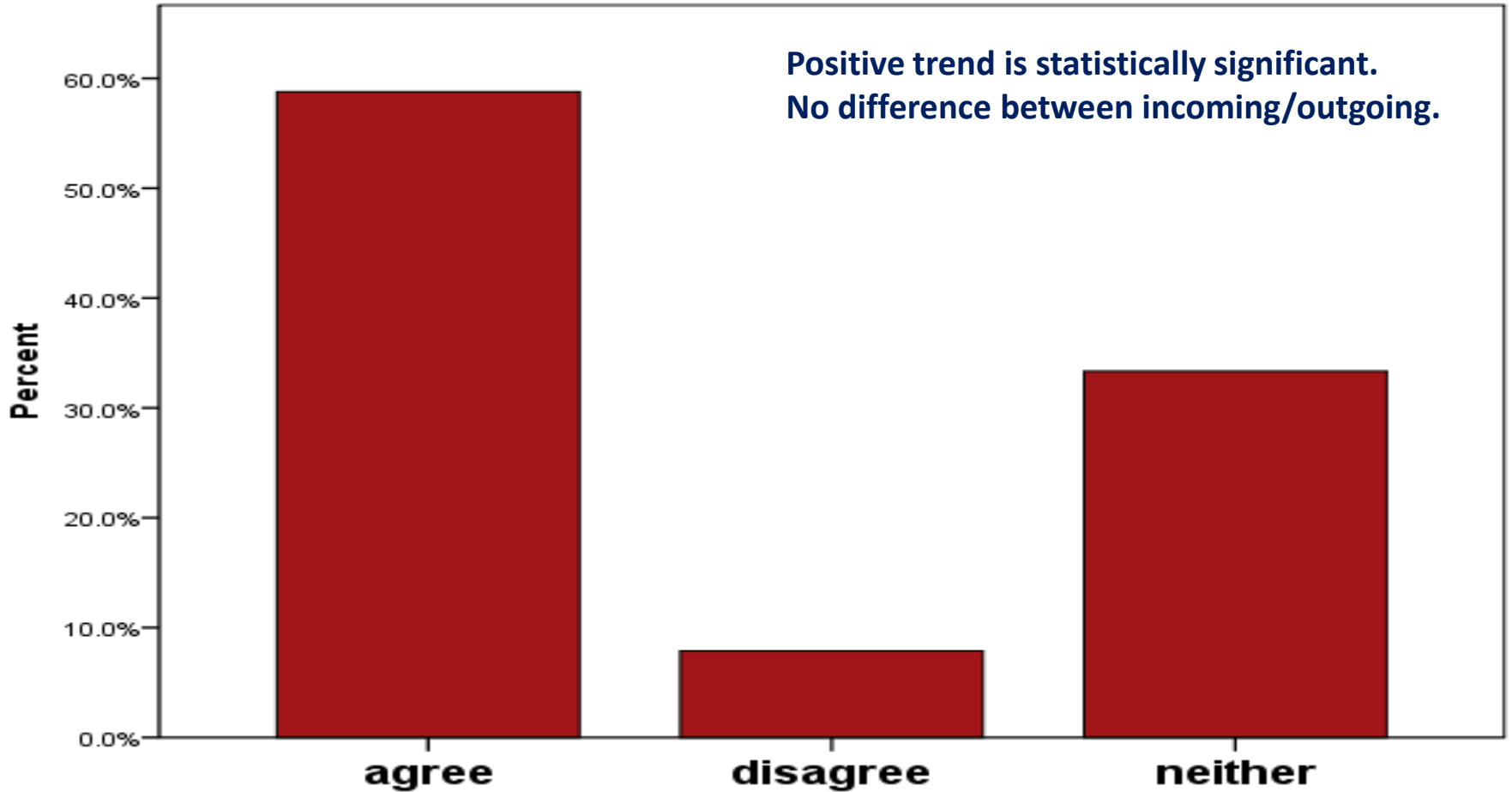
“The summary reflects what actually happened”



“The summary was easy to understand”



“The summary was helpful”



Comments from users

- We obtained comments for 138 scenarios out of 165 (83.6%)

Segmentation

- Comments were manually segmented.
 - 237 segments in all
- Content categories:
 - *Overall* (good/bad/neutral)
 - *Content* (good/unnecessary/missing...)
 - *Language* (good/poor)
- 3 independent annotators; good to moderate agreement using Kappa.

Example segmentation

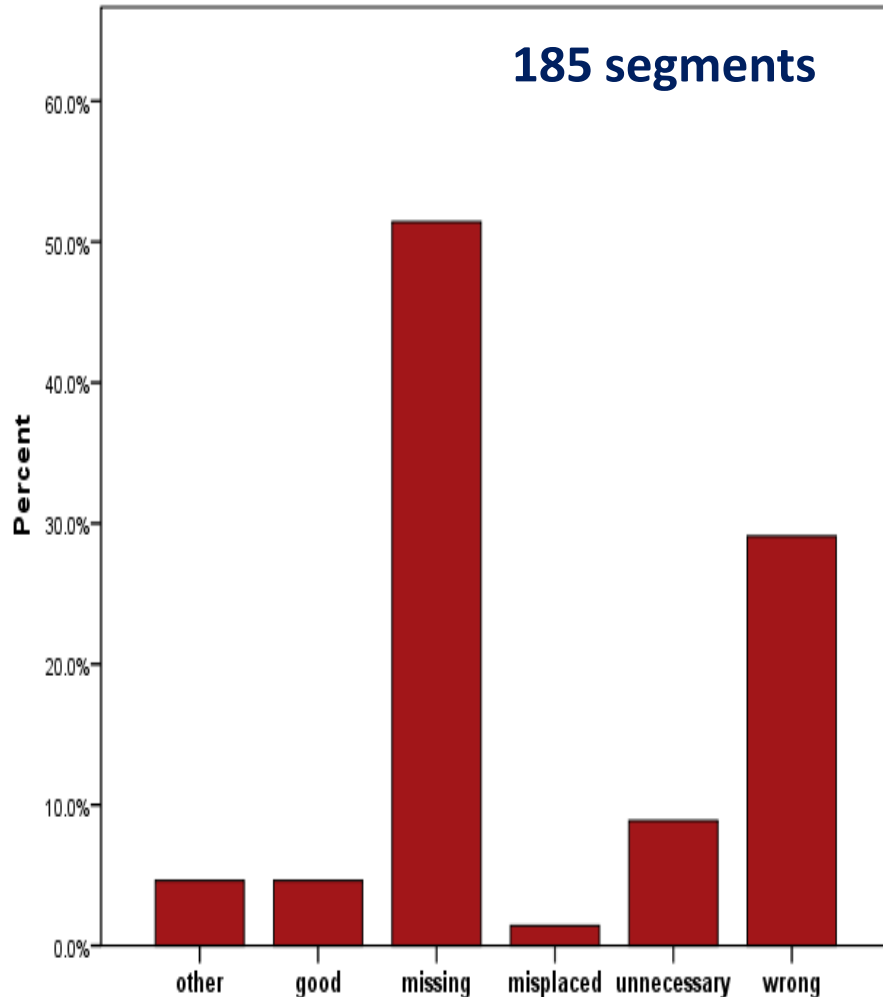
Example

- *Cardio - 'HR increased from 153 bpm to 154 bpm.' Not significant. Temp gap - ? Acceptable range. Would prefer a number*

Segmentation

- **Content:** *Cardio - 'HR increased from 153 bpm to 154 bpm.' Not significant.*
 - Category: *unnecessary*
- **Language:** *Temp gap - ? Acceptable range. Would prefer a number*
 - Category: *poor*

Summary content



- Mostly about missing or incorrect content.
 - *I cannot see how BT got the figure of 6.91 ml for urine output.*
- Sometimes directly relevant to BT's content selection & document planning:
 - *[BT has done a good job of reporting extubation and intubation,] though the listing of the dose of suxamethonium could have been with the note about intubation.*

Language

- Comments about wording or phrasing:
 - *Changes to vent settings is written in a confusing way.*
- Sometimes, BT-Nurse gets the pragmatics wrong:
 - *BT says: 'Between 22:30 and 02:30, FiO2 was raised from 40% to 51%'. When I read this phrase it makes me think that the baby spent those whole four hours at 51% oxygen.*
- 11 segments only.
- How do we interpret this?
 - Maybe language only draws attention to itself when it's really bad (or really good).
 - The absence of comments might mean that it was ok most of the time.

Summary

Summary

- Majority of users find summaries easy to understand, accurate and useful.
- Comments suggest that language is OK. Content issues could be resolved with more knowledge engineering.

Methodological pros:

- Real setup
- Real users
- No artificial tasks
- A solid test of system robustness

Methodological cons:

- Relatively uncontrolled (not an experimental design)
- Difficult to answer questions about decision-making accuracy
- Need to rely on user comments for evaluation of content selection and language generation.

Part 4

COMPARATIVE EVALUATION OF REFERRING EXPRESSION GENERATION

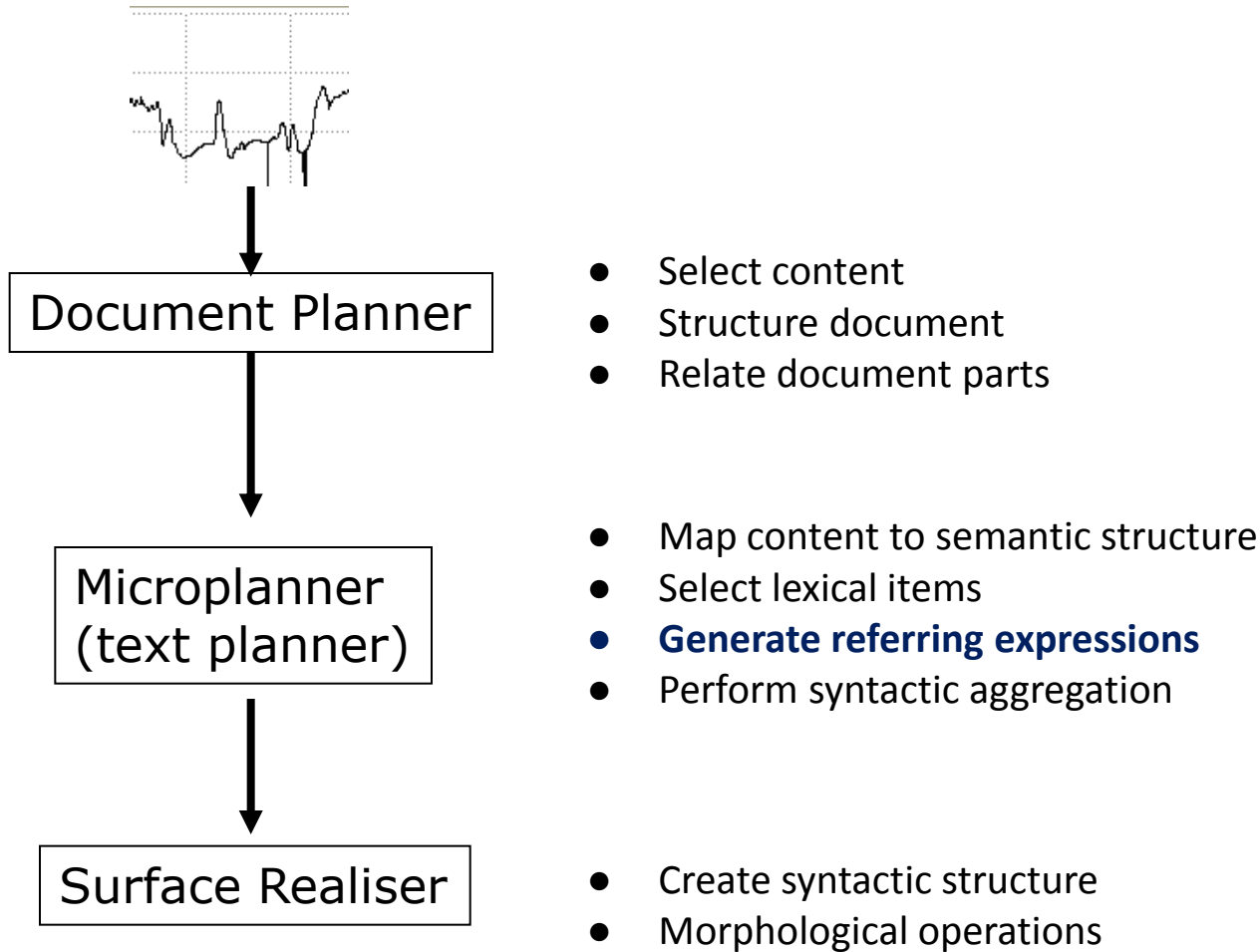
Comparative evaluation

- In many cases (as with BT), an evaluation depends on the nature of the system one has designed.
- However, in many areas of NLP, it is common to organise **shared tasks**:
 - A common input
 - A common task
 - Compare outputs in an evaluation
- The advantages are:
 - It's easier to see which solutions perform best and find the reasons why.
 - We have a lot of data for the same problem, and so can experiment with different evaluation methods and see whether they are comparable.

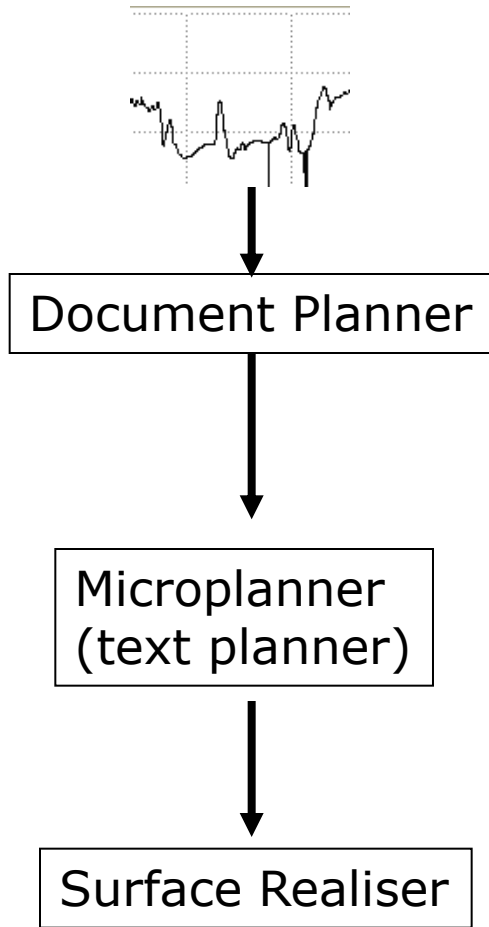
Case Study

- Generation Challenges has been organised since 2007.
- Wide range of different NLG tasks.
- We focus here on the TUNA tasks.
 - Comparison of algorithms for Generating Referring Expressions.
 - Based on data collected during the TUNA project.

NLG “consensus” architecture



NLG “consensus” architecture



- Select content
- Structure document
- Relate document parts

- Map content to semantic structure
- Select lexical items
- **Generate referring expressions**
- Perform syntactic aggregation

- Create syntactic structure
- Morphological operations

What does the system know about a discourse entity?
What properties should it use?
How should it be realised?

- *the bradycardia*
- *the bradycardia down to 69*
- *it*
- *the first bradycardia*

Generation of Referring Expressions

Input

- domain of relevant discourse entities
- a target referent

Output

- a noun phrase to identify that entity.

Subtasks

- Content determination
 - choosing what to say (the properties of the entity)
- Realisation
 - choosing how to say it
- An important component of many NLG systems.
- One of the most intensively studied tasks in NLG.

GRE Example



- the red chair facing back
- the large chair facing back
- the red chair
- the chair facing back
- it



Domain + referent

Distinguishing descriptions

GRE and comparative evaluation

Generation Challenges

- Series of shared tasks in areas of NLG
 - TUNA-REG Challenges organised as part of this, over three years (2007 – 2009)
 - Focus today: results from 2009
- GRE considered a very good candidate for the first shared tasks.
 - Significant agreement on task definition.
 - Data available: TUNA Corpus

Data & task

TUNA Corpus

- human-authored referring expressions of furniture or people;
- collected via an online elicitation experiment;
- human authors typed descriptions of referents in a visual context;
- referents belonged to **2 domains: furniture or people**.

Task definition

- Submitted systems needed to:
 - select the content of referring expressions
 - realise it as a string

Corpus data

Input



```
<DOMAIN>
  <ENTITY type="target">
    <ATTRIBUTE NAME="type" VALUE="person"/>
    <ATTRIBUTE NAME="hasHair" VALUE="0"/>
    <ATTRIBUTE NAME="hasBeard" VALUE="1"/>
    ...
  </ENTITY>
  <ENTITY type="distractor"> ... </ENTITY>
  ...
</DOMAIN>
```

Reference output

“the bald man with a beard”

```
<WORD-STRING>
  the bald man with a beard
</WORD-STRING>
```

Shared Task Setup

Original TUNA Corpus (singular section)

- 80% training data
- 20% development data

Test data

- 112 input DOMAINS
- 2 human outputs for each input DOMAIN
- equal no. of people and furniture cases

Participants

- 6 different systems in the 2009 edition
- many others in previous (2007-8) editions

Evaluation criteria in TUNA-REG

- Humanlikeness
- Adequacy/Clarity
- Fluency



Intrinsic methods:

Assess properties of systems in their own right

- Referential Clarity



Extrinsic method:

Assesses properties of systems in terms of its effect on human performance

Evaluation criteria

1. Humanlikeness

- compares system outputs to human outputs
- automatically computed

Measures of humanlikeness

1. String Edit (Levenshtein) Distance
 - number of insertions, deletions and substitutions to convert a peer description into the human description
2. BLEU-3
 - n-gram based string comparison
3. NIST-5
 - weighted version of BLEU, with more importance given to less frequent n-grams
4. Accuracy
 - proportion of outputs which are identical to the corresponding human description

Evaluation criteria

1. Humanlikeness

- compares system outputs to human outputs
- automatically computed

2. Adequacy

- judgement of adequacy of a description for the referent in its domain
- assessed by native speakers

Evaluation criteria

1. Humanlikeness

- compares system outputs to human outputs
- automatically computed

2. Adequacy

- judgement of adequacy of a description for the referent in its domain
- assessed by native speakers

3. Fluency

- judgement of fluency of description
- assessed by native speakers

Measures of adequacy and fluency

- Experiment with 8 linguistically aware native speakers
 - all postgraduate students in Language/Linguistics
- Participants shown:
 - system-generated or human-authored description
 - corresponding visual domain
- Answered two questions:
 - **How clear is this description?** (Is it clear what object it refers to?)
 - **How fluent is this description?** (Does it read well?)
- Ratings given using a slider (value between 1 and 100)
 - overcomes some of the objections to means comparison with interval scales

Experimental trial



Blue chair facing left.

Remember: the further to the left you place the slider, the more negative your judgement; the further to the right, the more positive your judgement.

How clear is this description? (Is it clear which object it refers to?)

How fluent is this description? (Does it read well?)

next

Evaluation criteria

1. Humanlikeness

- compares system outputs to human outputs
- automatically computed

2. Adequacy

- judgement of adequacy of a description for the referent in its domain
- assessed by native speakers

3. Fluency

- judgement of fluency of description
- assessed by native speakers

4. **Referential clarity (task-based)**

- speed and accuracy in an identification experiment
- performance on task as index of output quality

Measuring referential clarity

- Identification experiment with 16 participants

Procedure

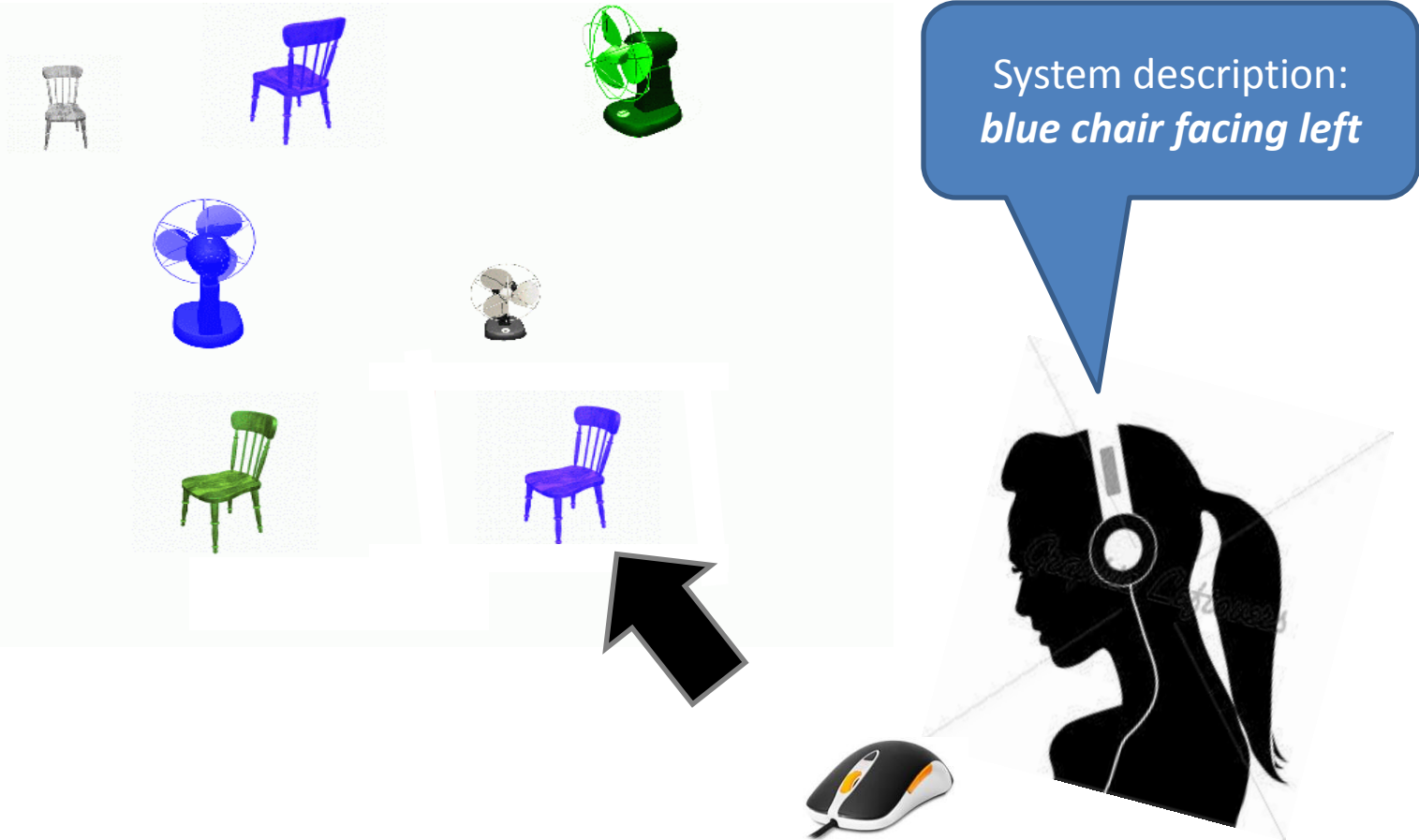
- participants shown a visual domain
- heard a description over headset produced using a TTS system
- clicked on the object identified

Measures

- **Identification speed (ms)**: how fast an object was identified
- **Identification accuracy (%)**: whether the correct (intended) object was identified

Referential clarity experimental setup

- Identification speed = speed of identification based on description
- Identification accuracy = error rate



Our main question

Are the different measures meaningfully related?

Do they tell us the same things about system quality?

Do they correlate?

Intrinsic human

	Fluency	Adequacy	Acc.	SE	BLEU	NIST	ID Acc.	ID Speed
Fluency	1	0.68						
Adequacy	0.68	1						
Accuracy								
SE								
BLEU								
NIST								
ID Acc.								
ID Speed								

Intrinsic human + intrinsic automatic

	Fluency	Adequacy	Acc.	SE	BLEU	NIST	ID Acc.	ID Speed
Fluency	1	0.68	0.85	-0.57	0.66	0.3		
Adequacy	0.68	1	0.83	-0.29	0.6	0.48		
Accuracy	0.85	0.83	1	-0.68	.86	0.49		
SE	-0.57	-0.29	-0.68	1	-0.75	-0.07		
BLEU	0.66	0.6	.86	-0.75	1	0.71		
NIST	0.3	0.48	0.49	-0.07	0.71	1		
ID Acc.								
ID Speed								

Intrinsic human + intrinsic automatic + extrinsic

	Fluency	Adequacy	Acc.	SE	BLEU	NIST	ID Acc.	ID Speed
Fluency	1	0.68	0.85	-0.57	0.66	0.3	0.5	-0.89
Adequacy	0.68	1	0.83	-0.29	0.6	0.48	0.95	-0.65
Accuracy	0.85	0.83	1	-0.68	.86	0.49	0.68	-0.79
SE	-0.57	-0.29	-0.68	1	-0.75	-0.07	-0.01	0.68
BLEU	0.66	0.6	.86	-0.75	1	0.71	0.49	-0.51
NIST	0.3	0.48	0.49	-0.07	0.71	1	0.6	0.06
ID Acc.	0.5	0.95	0.68	-0.01	0.49	0.6	1	-0.39
ID Speed	-0.89	-0.65	-0.79	0.68	-0.51	0.06	-0.39	1

Results: summary

What is the relationship between evaluation methods?

- no correlation between intrinsic automatic and extrinsic measures;
- intrinsic, human measures correlate better with extrinsic measures;
- intrinsic, automatic measures correlate only partially with intrinsic, human measures (Accuracy only)

Do these results generalise?

- Similar results found in two previous Shared Tasks (2007, 2008).
 - These included many more systems.

Some possible explanations

- Corpus contains a lot of individual variation (several authors).
- Not all referring expressions are of “high quality” (e.g. some are very telegraphic).
- BLEU/NIST more appropriate for long texts.

But still...

- What people do need not reflect what people like.
 - We should not always expect intrinsic automatic measures to correlate with judgements.
- What people do/like need not reflect what is effective.
 - We should not always expect intrinsic measures to correlate with task performance.

Some general conclusions

- Our results suggest that the relationship between intrinsic/corpus-based and extrinsic evaluation is problematic.
 - Referring expressions generation: 3 different sets of evaluation studies, over 3 years.
- In general, we should not rely on automatic techniques to make inferences about effectiveness of our NLG modules.

Data etc

- Generation Challenges data archive (includes TUNA data, plus data from other shared tasks in NLG):

<https://sites.google.com/site/genchalrepository/>

- TUNA Project (where the TUNA corpus was developed):

<http://www.abdn.ac.uk/ncs/computing/research/nlg/projects/previous/tuna/>