Evaluation of NLG Systems

Lecture 16
March 29, 2013

Johanna Moore
(slides adapted from Jon Oberlander)

School of informatics

Outline

- Distinguish types of NLG evaluation
- Automatic, intrinsic evaluation:
  - What's best?
  - Why are corpus-based gold standards problematic?
- Task-based, extrinsic evaluations
  - Are they too expensive?
- A way to feed back from task-based evaluations to help select best automatic, intrinsic metrics.

Some preliminaries - Belz 2009

- The user-oriented vs. developer-oriented distinction concerns evaluation purpose.
  - **Developer-oriented evaluations** focus on functionality … and seek to assess the quality of a system's (or component's) outputs.
  - **User-oriented evaluations** … look at a set of requirements (acceptable processing time, maintenance cost, etc.) of the user (embedding application or person) and assess how well different technological alternatives fulfill them.
- Another common distinction is about evaluation methods:
  - **Intrinsic evaluations** assess properties of systems in their own right, e.g., comparing their outputs to reference outputs in a corpus
  - **Extrinsic evaluations** assess the effect of a system on something that is external to it, for example, the effect on human performance at a given task or the value added to an application.
- Note also:
  - **Subjective** user evaluation (did you like the output/the system?)
  - **Objective** user evaluation (how fast/accurate are users on tasks?)

Adapted from Belz 2009

Intrinsic, developer evaluations hold sway

Most evaluation is of one of 3 basic intrinsic techniques

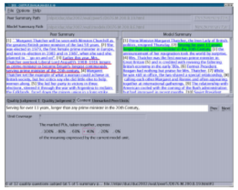
1. Assessment by trained assessors of the quality of system outputs according to different quality criteria, typically using rating scales
2. Automatic measurements of the degree of similarity between system outputs and reference outputs
   - E.g. BLEU and ROUGE
3. Human assessment of the degree of similarity between system outputs and reference outputs

What's missing is any form of extrinsic evaluation!

Adapted from Belz 2009

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Used for automatic, intrinsic evaluation of summarization systems



ROUGEs — N-gram co-occurrence metrics measuring content overlaps

Counts of N-gram overlaps between candidate and model summaries

$$ROUGE_n = \frac{\sum_{C \in \{Model\ Units\}} \sum_{n-gram \in C} Count_{match}(n-gram)}{\sum_{C \in \{Model\ Units\}} \sum_{n-gram \in C} Count(n-gram)}$$

Total number of n-grams in the model summary

Recall-based Metric!
(fixed-length summaries)

Chin-Yew Lin / MT Summit IX September 27, 2003, New Orleans, LA

But it's extrinsic evaluation that really matters

- "If we don't include *application purpose* in task definitions then not only do we not know which applications (if indeed any) systems are good for,
- we also don't know whether the task definition (including output representations) is appropriate for the application purpose we have in mind." (p. 113)

Adapted from Belz 2009

Why we often settle for less …

- For NL understanding, there is usually a single target output.
  - But for generation, with multiple outputs, similarity to reference texts matters
- Metrics like BLEU and ROUGE are only "surrogate measures"
  - We test them via their "correlation with human ratings of quality,
    - using Pearson's product-moment correlation coefficient or Spearman's rank-order correlation coefficient"
    - Stronger the correlation, the better the metric
  - We don't then test the human ratings
  - "If human judgment says a system is good, then if an automatic measure says the system is good, it simply confirms human judgment; if the automatic measure says the system is bad, then the measure is a bad one"
  - But if intrinsic conflicts with extrinsic, should be worried

Adapted from Belz 2009

Intrinsic vs extrinsic - reasons to be concerned

- Law et al. (2005)
  - Compared graphical representations of medical data with textual descriptions of same data
    - in intrinsic assessments doctors rated the graphs more highly than the texts
    - ***but*** in extrinsic diagnostic performance test they performed better with the texts than the graphs
- Engelhardt, Bailey, and Ferreira (2006)
  - subjects rated over-descriptions as highly as concise descriptions,
  - ***but performed worse*** at a visual identification task with over-descriptions than with concise descriptions
- Miyao et al. (2008)
  - Performed evaluation of 8 parsers used in Biomedical IR system
  - Effect parsers had on IR quality showed different ranking than intrinsic evaluation using F-scores

Adapted from Belz 2009

Further reasons for concern

- "Stable averages of human quality judgments, let alone high levels of agreement, are hard to achieve"
  - Recall SPaRKy
- Does a human top line always mean machines must perform more poorly?
  - "In NLG, domain experts have been shown to prefer system-generated language to alternatives produced by human experts" (Belz & Reiter, EACL 2006)
- "The explanation routinely given for not carrying out extrinsic evaluations is that they are too time-consuming and expensive."
  - Later on, we will question the validity of that position.

Comparative evaluation

- Evaluation often depends on the nature of the system one has designed. Hard to compare results if different task, inputs, expected outputs, etc.
- 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 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: GRE and comparative evaluation

Generation Challenges

- Series of shared tasks in wide range of NLG tasks
  - TUNA-REG Challenges: comparison of algorithms for Generating Referring Expressions)GRE)
    - over three years (2007 – 2009)
    - focus today: results from 2009
  - GIVE Challenge: Giving Instructions in a Virtual Environment
- GRE considered a very good candidate for the first shared tasks.
  - Significant agreement on task definition.
  - Data available: TUNA Corpus

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

Adapted from slide by Gatt

Data & task

TUNA Corpus

- human-authored referring expressions of furniture or people
  - collected via an online elicitation experiment using University of Zurich Web Experimentation List website
  - human authors presented with scene and typed descriptions of referents
- paired with representation of entities and attributes

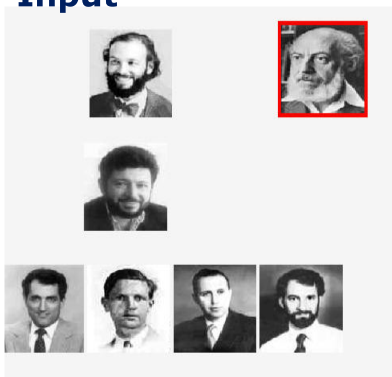
Task definition

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

Adapted from slide by Gatt

Data from People Corpus

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>
```

Adapted from slide by Gatt

Shared Task Setup

Original TUNA Corpus

- 80% training data
- 20% development data

	Furniture	People	All
Training	319	274	593
Development	80	68	148
Test	56	56	112
All	455	398	853

Test data

- 112 input domains: entities & attributes
- 2 human outputs for each input domain
- equal number of people and furniture cases

Participants

- 6 different systems in the 2009 edition

Teams

- IS:
  - extended full-brevity algorithm which uses a nearest neighbour technique to select the attribute set (AS) most similar to a given writer's previous ASs
- GRAPH:
  - existing graph-based attribute selection component, which represents a domain as a weighted graph, and uses a cost function for attributes. Team developed a new realiser which uses a set of templates derived from the descriptions in the TUNA corpus.
- NIL-UCM:
  - three systems submitted by this group use a standard evolutionary algorithm for attribute selection
- USP:
  - system USP-EACH, is a frequency-based greedy attribute selection strategy

Adapted from Gatt, Belz and Kow 2009

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 effect on human performance

Adapted from slide by Gatt

Evaluation criteria: Human intrinsic

1. **Humanlikeness**
   - compares system outputs to human outputs
   - automatically computed

Adapted from slide by Gatt

Computing 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 that are identical to the corresponding human description

Adapted from slide by Gatt

Evaluation criteria: Human Intrinsic

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

Adapted from slide by Gatt

Evaluation criteria: Human Intrinsic

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

Adapted from slide by Gatt

Measures of adequacy and fluency: Human Intrinsic

- 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:

  Q1: How clear is this description? Try to imagine someone who could see the same grid with the same pictures, but didn't know which of the pictures was the target. How easily would they be able to find it, based on the phrase given?

  Q2: How fluent is this description? Here your task is to judge how well the phrase reads. Is it good, clear English?"
- Ratings given using a slider (value between 1 and 100)
  – overcomes some of the objections to means comparison with interval scales

Adapted from slide by Gatt

Experimental trial

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

Measuring referential clarity

Identification experiment with 16 participants

Procedure:

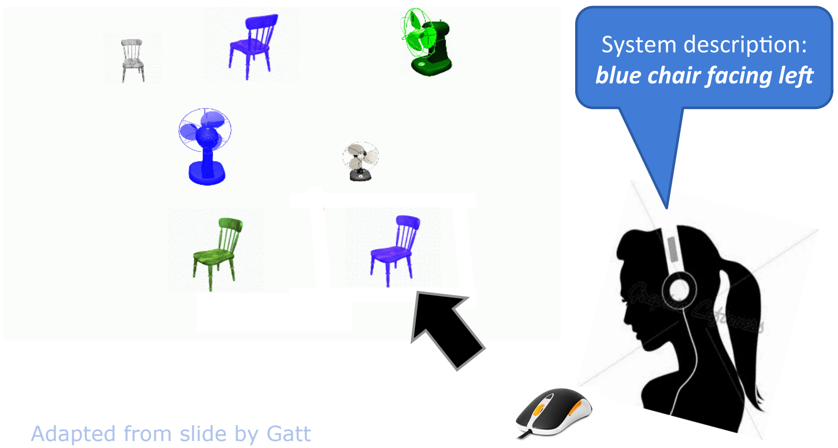
- participants shown a visual domain
- heard a description over headset produced using a TTS system
- clicked on the object identified

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 questions

- Are the different measures meaningfully related?
- Do they tell us the same things about system quality?
- Do they correlate with one another?

Evaluation criterion	*Type of evaluation*	*Evaluation technique*
Humanlikeness	Intrinsic/automatic	Accuracy, String-edit distance, BLEU-3, NIST
Adequacy/clarity	Intrinsic/human	Judgment of adequacy as rated by native speakers
Fluency	Intrinsic/human	Judgment of fluency as rated by native speakers
Referential clarity	Extrinsic/human	Speed and accuracy in identification experiment

Results - ranked by String Edit Distance

	All test data			
	Acc	SE	BLEU	NIST
GRAPH	12.50	6.41	0.47	2.57
IS-FP-GT	3.57	6.74	0.28	0.75
NIL-UCM-EvoTAP	6.25	7.28	0.26	0.90
USP-EACH	7.14	7.59	0.27	1.33
NIL-UCM-ValuesCBR	2.68	7.71	0.27	1.69
NIL-UCM-EvoCBR	2.68	8.02	0.26	1.97
HUMAN-2	2.68	9.68	0.12	1.78
HUMAN-1	2.68	9.68	0.12	1.68

One way ANOVA for SE scores:

- All systems significantly better than human-authored
- GRAPH better than NIL-ICM-EvoCBR

Adapted from Gatt, Belz & Kow 2009

Results - ranked by Adequacy

	All test data			
	Adequacy		Fluency	
	Mean	SD	Mean	SD
GRAPH	84.11	21.07	85.81	17.52
USP-EACH	77.72	28.33	84.20	20.27
NIL-UCM-EvoTAP	76.16	28.34	61.95	26.13
HUMAN-2	74.63	34.77	73.38	27.63
NIL-UCM-ValuesCBR	72.34	33.93	59.41	33.94
HUMAN-1	70.38	34.92	71.52	30.79
NIL-UCM-EvoCBR	63.65	37.19	55.38	35.32
IS-FP-GT	59.46	40.94	66.21	30.97

Systems which do not share a letter are significantly different at α = .05

Adequacy					Fluency					
GRAPH	A				GRAPH	A				
USP-EACH	A	B			USP-EACH	A	B			
NIL-UCM-EvoTAP	A	B			HUMAN-2		B	C		
HUMAN-2	A	B	C		HUMAN-1			C	D	
NIL-UCM-ValuesCBR	A	B	C		IS-FP-GT			C	D	E
HUMAN-1		B	C	D	NIL-UCM-EvoTAP				D	E
NIL-UCM-EvoCBR			C	D	NIL-UCM-ValuesCBR					E
IS-FP-GT				D	NIL-UCM-EvoCBR					E

Results – identification accuracy and speed

	All test data		
	ID acc.	ID. speed	
	%	Mean	SD
GRAPH	0.96	3069.16	878.89
HUMAN-1	0.91	3517.58	1028.83
USP-EACH	0.90	3067.16	821.00
NIL-UCM-EvoTAP	0.88	3159.41	910.65
NIL-UCM-ValuesCBR	0.87	3262.53	974.55
HUMAN-2	0.83	3463.88	1001.29
NIL-UCM-EvoCBR	0.81	3362.22	892.45
IS-FP-GT	0.68	3167.11	964.45

USP-EACH	A	
GRAPH	A	
NIL-UCM-EvoTAP	A	B
IS-FP-GT	A	B
NIL-UCM-ValuesCBR	A	B
NIL-UCM-EvoCBR	A	B
HUMAN-2		B
HUMAN-1		B

Identification Speed:
Systems that do not share a letter are significantly different at α = .05

Adapted from Gatt, Belz and Kow 2009

Our main questions

Are the different measures meaningfully related?

- Do they tell us the same things about system quality?
- Do they correlate with one another?

Evaluation criterion	*Type of evaluation*	*Evaluation technique*
Humanlikeness	Intrinsic/automatic	Accuracy, String-edit distance, BLEU-3, NIST
Adequacy/clarity	Intrinsic/human	Judgment of adequacy as rated by native speakers
Fluency	Intrinsic/human	Judgment of fluency as rated by native speakers
Referential clarity	Extrinsic/human	Speed and accuracy in identification experiment

Adapted from slide by Gatt

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 (IH) + Intrinsic Automatic (IA)

	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 (EX)

	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

How do the various measures correlate? Summary

- EX id-accuracy significantly correlated with IH adequacy (+)
- EX id-speed significantly correlated with IH fluency (-)
- IA accuracy significantly correlated with
  - IH fluency (+) and IH adequacy (+)
  - IA BLEU (+)

	Human-assessed, intrinsic		Extrinsic		Auto-assessed, intrinsic			
	Fluency	Adequacy	ID Acc.	ID Speed	Acc.	SE	BLEU	NIST
Fluency	1	0.68	0.50	-0.89*	.85*	-0.57	0.66	0.30
Adequacy	0.68	1	0.95**	-0.65	.83*	-0.29	0.60	0.48
Identification Accuracy	0.50	0.95**	1	-0.39	0.68	-0.01	0.49	0.60
Identification Speed	0.89*	-0.65	-0.39	1	-0.79	0.68	-0.51	0.06
Accuracy	0.85*	0.83*	0.68	-0.79	1.00	-0.68	.859*	0.49
SE	-0.57	-0.29	-0.01	0.68	-0.68	1	-0.75	-0.07
BLEU	0.66	0.60	0.49	-0.51	.86*	-0.75	1	0.71
NIST	0.30	0.48	0.60	0.06	0.49	-0.07	0.71	1

Adapted from Gatt, Belz and Kow 2009

Are corpus-based intrinsic measures OK?

- "When automatically evaluating generated output, the goal is to find metrics that can easily be computed and that can also be shown to correlate with human judgments of quality.
- Many automated generation evaluations measure the similarity between the generated output and a corpus of gold-standard target outputs, often using measures such as precision and recall.
- Such measures of corpus similarity are straightforward to compute and easy to interpret; however, they are not always appropriate for generation systems.
- Several recent studies … have shown that strict corpus-similarity measures tend to favour repetitive generation strategies that do not diverge much, on average, from the corpus data, while human judges often prefer output with more variety."

Adapted from Foster 2008

(a) Right turn + brow raise (b) Left lean + brow lower

(a) Neutral (b) Right turn + brow raise (c) Left lean + brow lower

Animating an embodied conversational agent (ECA)

- Most common display used by the speaker was a downward nod
- User-preferences had the single largest differential effect on the displays used
  - When speaker described features of the design that user was expected to like, he was more likely to turn to the right and raise eyebrows
  - on features that user was expected to dislike he was more likely to lean left, lower eyebrows, and narrow eyes

From Foster 2008

Relating these measures back to human judgments

- Devised 3 algorithms for controlling ECA
- Collected users preference judgments for alternatives (Foster and Oberlander 2007)
- None of the corpus-reproduction metrics had any relationship to the users' preferences
- Number and diversity of displays per sentence contributed much more strongly to human judgments

Don't use similarity to corpus as your gold standard!

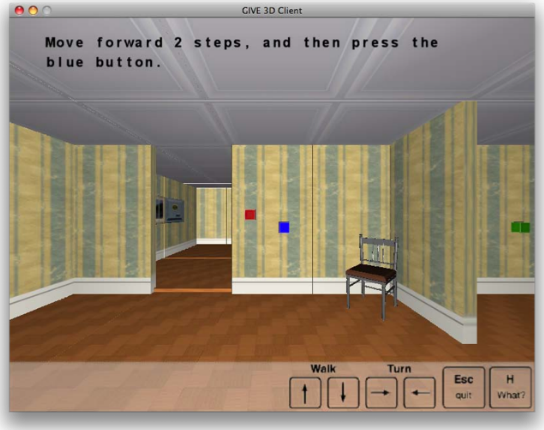
Adapted from Foster 2008

Is Extrinsic Evaluation Always Too Expensive?

- Not necessarily
- Crowd sourcing using the web
  - Amazon Mechanical Turk
  - Generating Instructions in Virtual Environments (GIVE) Challenge

Shared tasks in NLG - GIVE - Koller et al. 2009

- Subjects solve a treasure hunt in a virtual 3D world they have not seen before
- System has a complete symbolic representation of the virtual world
- Challenge for NLG system is to generate, in real time, natural-language instructions to guide user to successful completion of their task



Adapted from Koller et al. 2009

The GIVE software architecture

1. Client displays 3D world to users and allows them to interact with it
2. NLG servers generate the natural-language instructions
3. Matchmaker establishes connections between clients and NLG servers

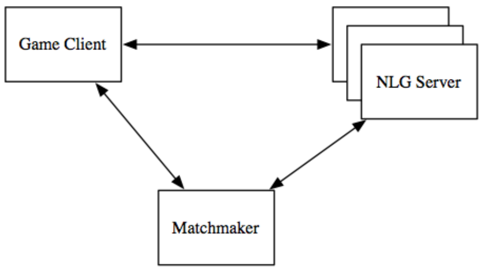


Figure 2: The GIVE architecture.

Adapted from Koller et al. 2009

Is Extrinsic Evaluation Always Too Expensive?

- Not necessarily
- Crowd sourcing using the web
  - Amazon Mechanical Turk
- Generating Instructions in Virtual Environments (GIVE) Challenge

In three months, collected 1143 valid games

JAST - evaluating a Joint Action robot

Adapted from Guiliani et al. 2010

Experimented with dialogue strategies and with reference generation

1. System First we will build a windmill.

2. User Okay.

3. User {*picks up a yellow cube, unnecessary piece for a windmill*}

4. System You don't need a yellow cube to build a windmill.

5. System To build a windmill, you first need to build a tower.

6. System *[picking up and holding out red cube]* To build the tower, insert the green bolt through the end of this red cube and screw it into the blue cube.

7. User *[takes cube, performs action]* Okay.

Adapted from Guiliani et al. In Submission

Task-based evaluation: Subjective and objective results

Table 1: Overall usability results

	Constant	Adaptive	M-W
Intell.	79.0 (15.6)	74.9 (12.7)	$p = 0.19$, n.s.
Task	72.7 (10.4)	71.1 (8.3)	$p = 0.69$, n.s.
Feeling	66.9 (15.9)	66.8 (14.2)	$p = 0.51$, n.s.
Conv.	66.1 (13.6)	75.2 (10.7)	$p = 0.036$, sig.
Overall	72.1 (11.2)	71.8 (9.1)	$p = 0.68$, n.s.

Table 2: Objective results (all differences n.s.)

Measure	Constant	Adaptive	M-W
Duration (s.)	404.3 (62.8)	410.5 (94.6)	$p = 0.90$
Duration (turns)	29.8 (5.02)	31.2 (5.57)	$p = 0.44$
Rep requests	0.26 (0.45)	0.32 (0.78)	$p = 0.68$
Explanations	2.21 (0.63)	2.41 (0.80)	$p = 0.44$
Successful trials	1.58 (0.61)	1.55 (0.74)	$p = 0.93$

Adapted from Guiliani et al. In Submission

Connecting task with available intrinsic metrics - PARADISE

- The PARADISE evaluation framework (Walker et al., 2000) explores the relationship between the subjective and objective factors.
- PARADISE uses stepwise multiple linear regression to predict subjective user satisfaction
- based on measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency, resulting in a predictor function

Table 3: PARADISE predictor functions for each category on the usability questionnaire

Measure	Function	R^2	Significance
Intelligence	$76.8 + 7.00 * \mathcal{N}(\text{Correct}) - 5.51 * \mathcal{N}(\text{Repeats})$	0.39	Correct: $p < 0.001$, Repeats: $p < 0.005$
Task	$72.4 + 3.54 * \mathcal{N}(\text{Correct}) - 3.45 * \mathcal{N}(\text{Repeats}) - 2.17 * \mathcal{N}(\text{Explain})$	0.43	Correct: $p < 0.005$, Repeats: $p < 0.01$, Explain: $p \approx 0.10$
Feeling	$66.9 - 6.54 * \mathcal{N}(\text{Repeats}) + 4.28 * \mathcal{N}(\text{Seconds})$	0.09	Repeats: $p < 0.05$, Seconds: $p \approx 0.12$
Conversation	$71.0 + 5.28 * \mathcal{N}(\text{Correct}) - 3.08 * \mathcal{N}(\text{Repeats})$	0.20	Correct: $p < 0.01$, Repeats: $p \approx 0.10$
Overall	$72.0 + 4.80 * \mathcal{N}(\text{Correct}) - 4.27 * \mathcal{N}(\text{Repeats})$	0.40	Correct: $p < 0.001$, Repeats: $p < 0.005$

Summary

- Much work has focused on automatic, intrinsic evaluation
- Some metrics are related to human, intrinsic evaluations.
  - But they're still only a surrogate for extrinsic evaluation!
- Temptation to use automatic corpus-based metrics should be resisted - some other automatic metrics may be superior, especially when variation is valued.
- Task-based, extrinsic evaluations are the best, and are not as expensive as sometimes been claimed.
- PARADISE can allow findings from task-based evaluations to feed back into appropriate engineering choices and selection of appropriate automatic, intrinsic metrics.

References

- Belz, A. (2009) That's nice ... what can you do with it? *Computational Linguistics*, 35, 111-118.
- Byron, D., Koller, A., Striegnitz, K., Cassell, J., Dale, R., Moore, J. and Oberlander, J. (2009) Report on the First NLG Challenge on Generating Instructions in Virtual Environments (GIVE). In *Proc of the 12th European Workshop on Natural Language Generation*.
- Foster, M.E. (2008). Automated metrics that agree with human judgments on generated output for an embodied conversational agent. *In Proc of INLG 2008*.
- Foster, M.E. and Oberlander, J. (2007) Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *International Journal of Language Resources and Evaluation*, 41:305-323
- A. Gatt, A. Belz and E. Kow (2009). The TUNA-REG Challenge 2009: Overview and evaluation results. *Proc. of the 12th European Workshop on Natural Language Generation* (ENLG-09).
- Giuliani, M. et al. (2010) Situated Reference in a Hybrid Human-Robot Interaction System. In *Proc of INLG 2010*.
- Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Dalzel-Job, S., Oberlander, J. and Moore, J. (2009) Validating the web-based evaluation of NLG systems. *In Proc of ACL-47*.
- M. Walker, C. Kamm, and D. Litman (2000). Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6:363-377.