Evaluating Dialogue Systems

Kolja Kirsch

Uni Hamburg

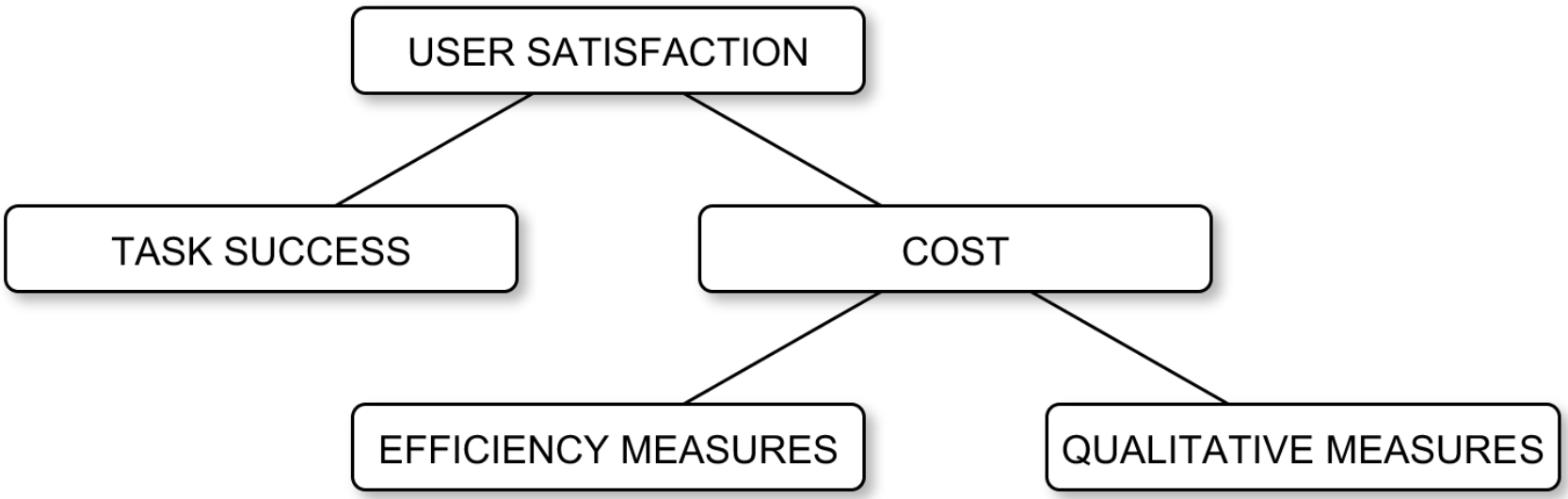
0kirsch@informatik.uni-hamburg.de

June 22, 2016

Why?

- How well does a system meet (different kinds of) goals?
- Compare systems (and tasks)
- Find possible improvements

PARADISE

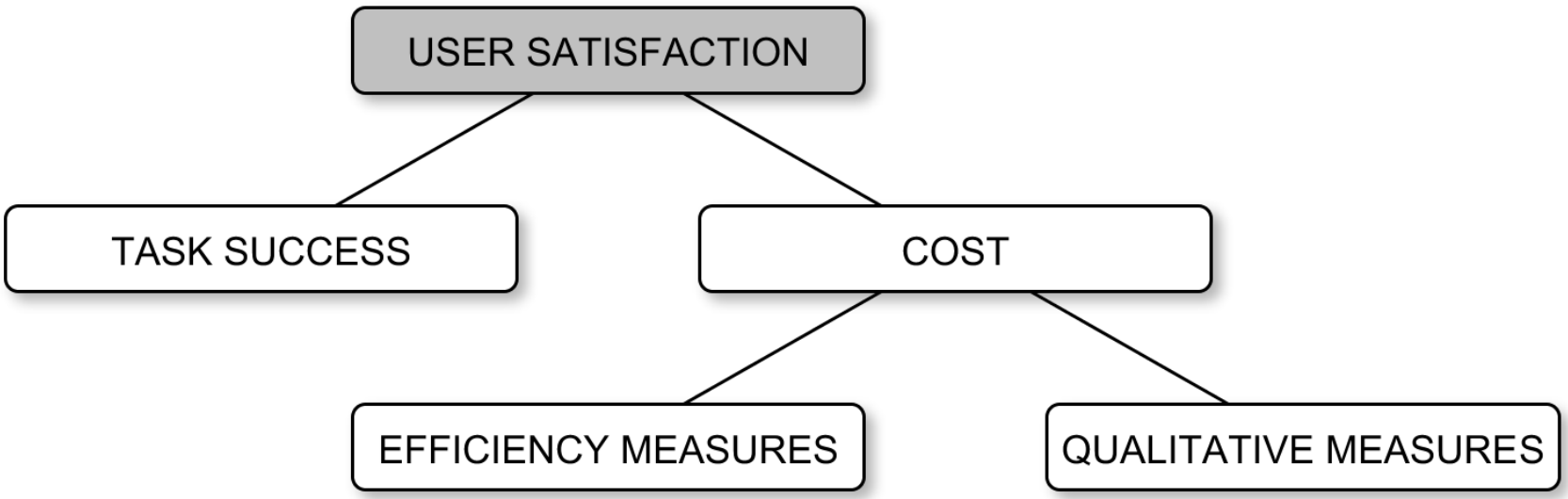

USER SATISFACTION
TASK SUCCESS
COST
EFFICIENCY MEASURES
QUALITATIVE MEASURES

PARADISE

A1:	Hello, This is Train Enquiry Service. Please speak after the tone. Which information do you need?
U1:	I want to go from Torino to Milano.
A2:	Do you want to go from Trento to Milano? Yes or No?
U2:	No.
A3:	Do you want to leave from Trento? Yes or No?
U3:	No.
A4:	Where do you want to leave from?
U4:	I want to leave from Torino.
A5:	Do you want to leave from Torino? Yes or No?
U5:	Yes.
A6:	Do you want to go to Milano?
U6:	Yes.
A7:	At which time do you want to leave?
U7:	I want to travel in the evening.
A8:	Do you want to leave between 6 and 9 p.m.? Yes or No?
U8:	Yes.
A9:	There is a train leaving at 8:00 p.m.
	. . .

B1:	Hello, This is Train Enquiry Service. Please speak after the tone. Which information do you need?
U1:	I want to travel from Torino to Milano.
B2:	At which time do you want to leave from Merano to Milano?
U2:	No, I want to leave from Torino in the evening.
B3:	Do you want to leave from Torino between 6 and 11 p.m? Please answer Yes or No.
U3:	Yes.
B4:	A train leaves at 8 p.m.
	. . .

PARADISE


USER SATISFACTION
TASK SUCCESS
COST
EFFICIENCY MEASURES
QUALITATIVE MEASURES

PARADISE: User satisfaction

I am satisfied with the performance of the system

Strongly disagree 1–2–3–4–5–6–7 Strongly agree

It is simple to use

Strongly disagree 1–2–3–4–5–6–7 Strongly agree

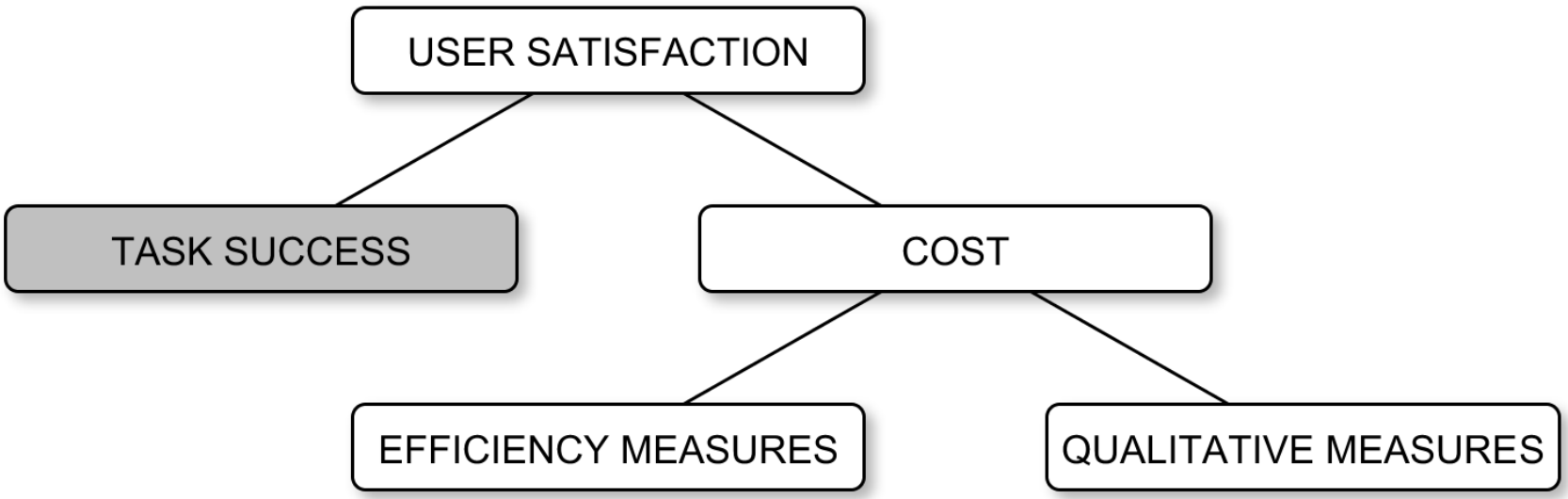
It is fun to use

Strongly disagree 1–2–3–4–5–6–7 Strongly agree

It does what I expect it to do

Strongly disagree 1–2–3–4–5–6–7 Strongly agree

PARADISE


USER SATISFACTION
TASK SUCCESS
COST
EFFICIENCY MEASURES
QUALITATIVE MEASURES

PARADISE: Task success

attribute	possible values	information flow
depart-city (DC)	Milano, Roma, Torino, Trento	to agent
arrival-city (AC)	Milano, Roma, Torino, Trento	to agent
depart-range (DR)	morning,evening	to agent
depart-time (DT)	6am,8am,6pm,8pm	to user

attribute	actual value
depart-city	Torino
arrival-city	Milano
depart-range	evening
depart-time	8pm

- Kappa coefficient (Carletta 1996)

PARADISE: Task success

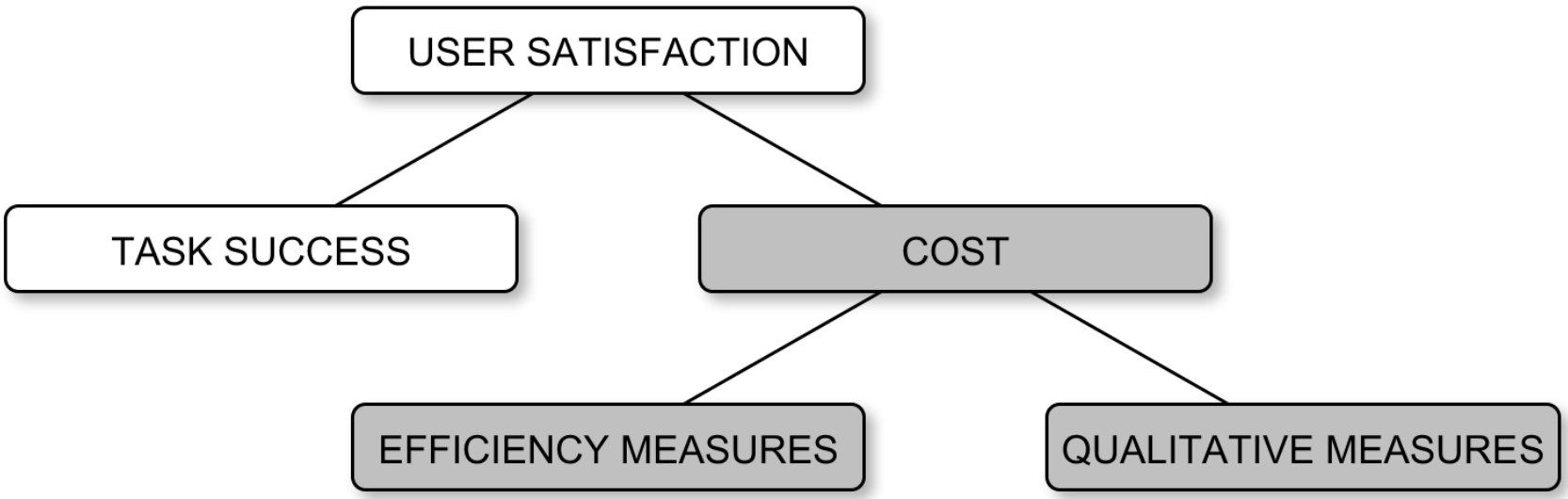
	KEY													
	DEPART-CITY				ARRIVAL-CITY				DEPART-RANGE		DEPART-TIME			
DATA	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14
v1	**16**		1		4				3	2				
v2	1	**20**	1			3								
v3	5	1	**9**	4	2		4	2						
v4	1	2	6	**6**			2	3						
v5	4				**15**				2	3				
v6	1	6				**19**								
v7			5	2	1	1	**15**	4						
v8		1	3	3	1	2	9	**11**						
v9	2				2				**39**	10				
v10									6	**35**				
v11											**20**	5	5	4
v12												**10**	5	5
v13											5	5	**10**	5
v14												5	5	**11**
sum	30	30	25	15	25	25	30	20	50	50	25	25	25	25

Figure: Confusion matrix for Agent B

PARADISE: Task success

- Actual agreement: $P(A) = \frac{\sum_{i=1}^{n} M(i,i)}{T}$
- Expected agreement: $P(E) = \sum_{i=1}^{n} M((\frac{t_i}{T})^2$
- Kappa coefficient: $\kappa = \frac{P(A)-P(E)}{1-P(E)}$

PARADISE: Cost



PARADISE: Cost

Efficiency measures

- Number of utterances
- Dialogue time
- ...

Qualitative measures

- Response delay
- Number of repairs
- ...

What else could you measure?

PARADISE: Performance

$$\text{Performance} = (\alpha * N(\kappa)) - \sum_{i=1}^{n} w_i * N(c_i)$$

user	agent	US	κ	c_1 (#utt)	c_2 (#rep)
1	A	1	1	46	30
2	A	2	1	50	30
3	A	2	1	52	30
4	A	3	1	40	20
5	A	4	1	23	10
6	A	2	1	50	36
7	A	1	0.46	75	30
8	A	1	0.19	60	30
9	B	6	1	8	0
10	B	5	1	15	1
11	B	6	1	10	0.5
12	B	5	1	20	3
13	B	1	0.19	45	18
14	B	1	0.46	50	22
15	B	2	0.19	34	18
16	B	2	0.46	40	18
Mean(A)	A	2	0.83	49.5	27
Mean(B)	B	3.5	0.66	27.8	10.1
Mean	NA	2.75	0.75	38.6	18.5

PARADISE

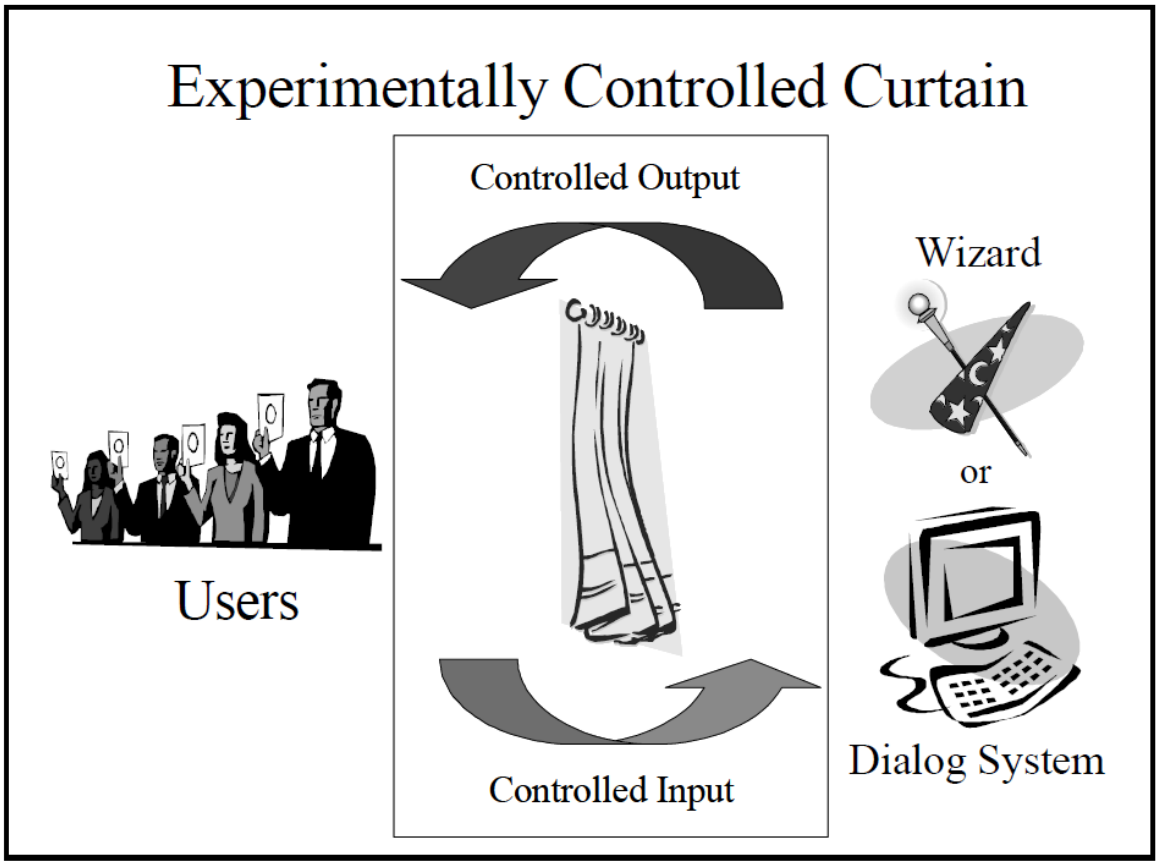
Is this really PARADISE?

PARADISE

- Not all factors have to be significant
- How much of the variance can be explained? (R^2)
- Significance of the performance
- Hidden variables
- What does the performance score mean?
- Why linear regression?

Hidden variables & interpreting scores

- Gold standard: Human conversation
- baseline for comparison



Hidden variables & interpreting scores

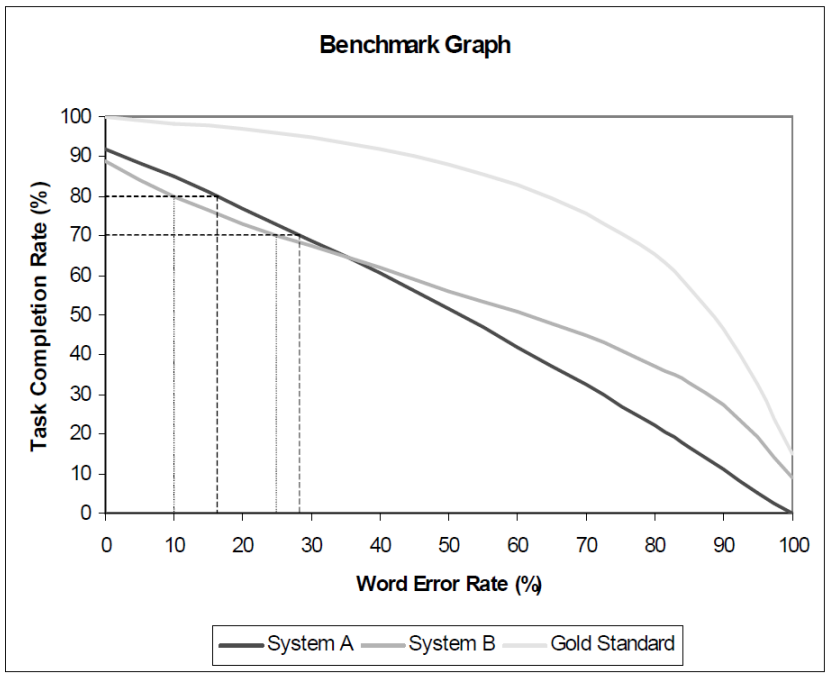


Figure 2. Comparison of two dialog systems with respect to the gold standard.

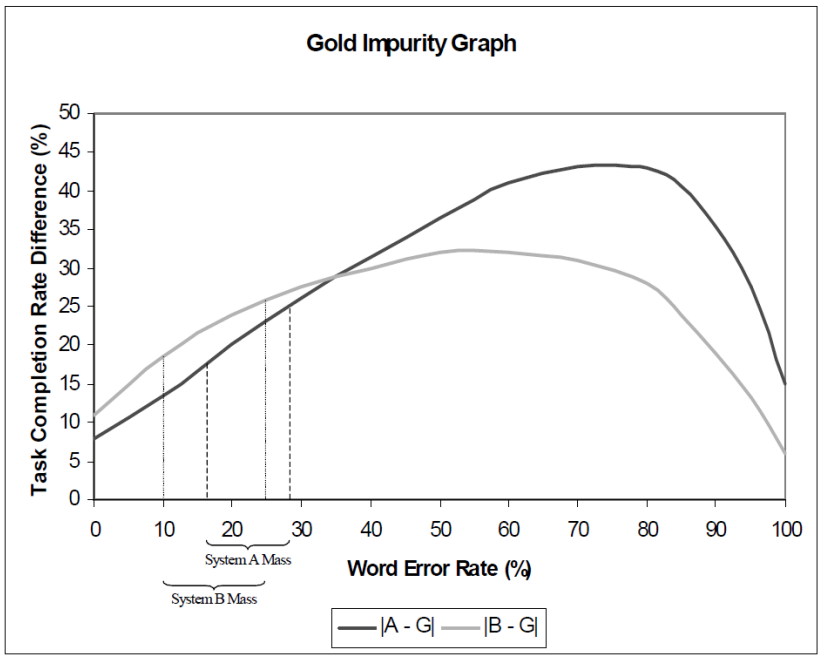
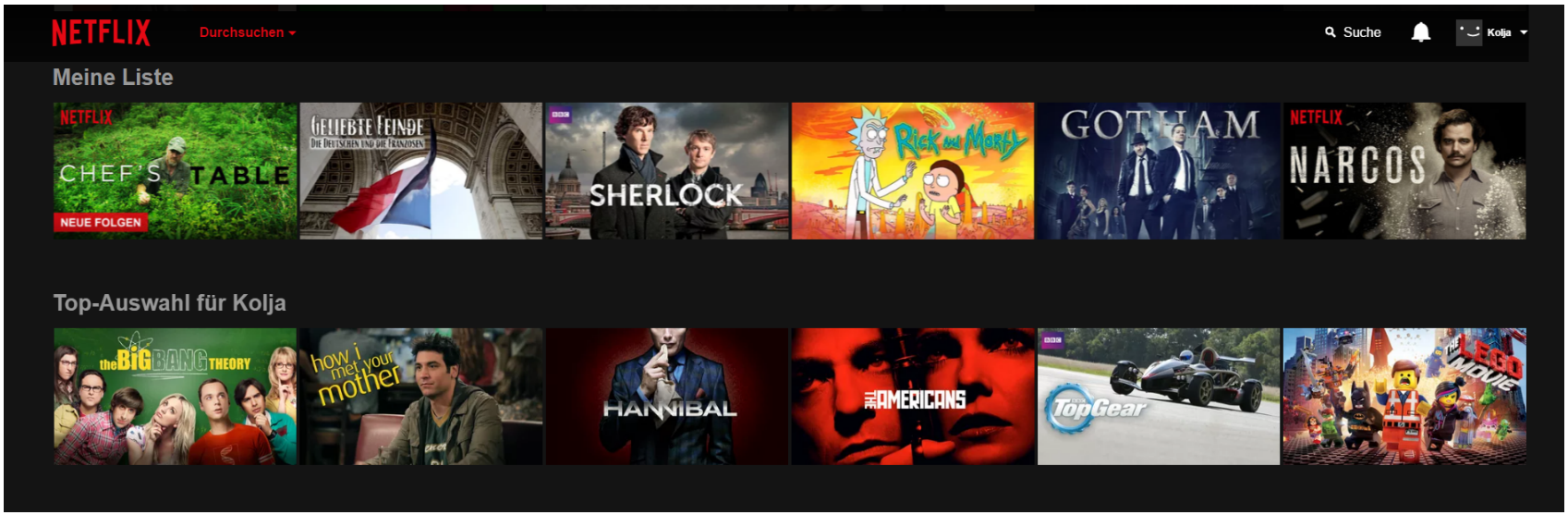


Figure 3. Distance in performance of the two systems from the gold standard.

Colaborative filtering



Colaborative filtering

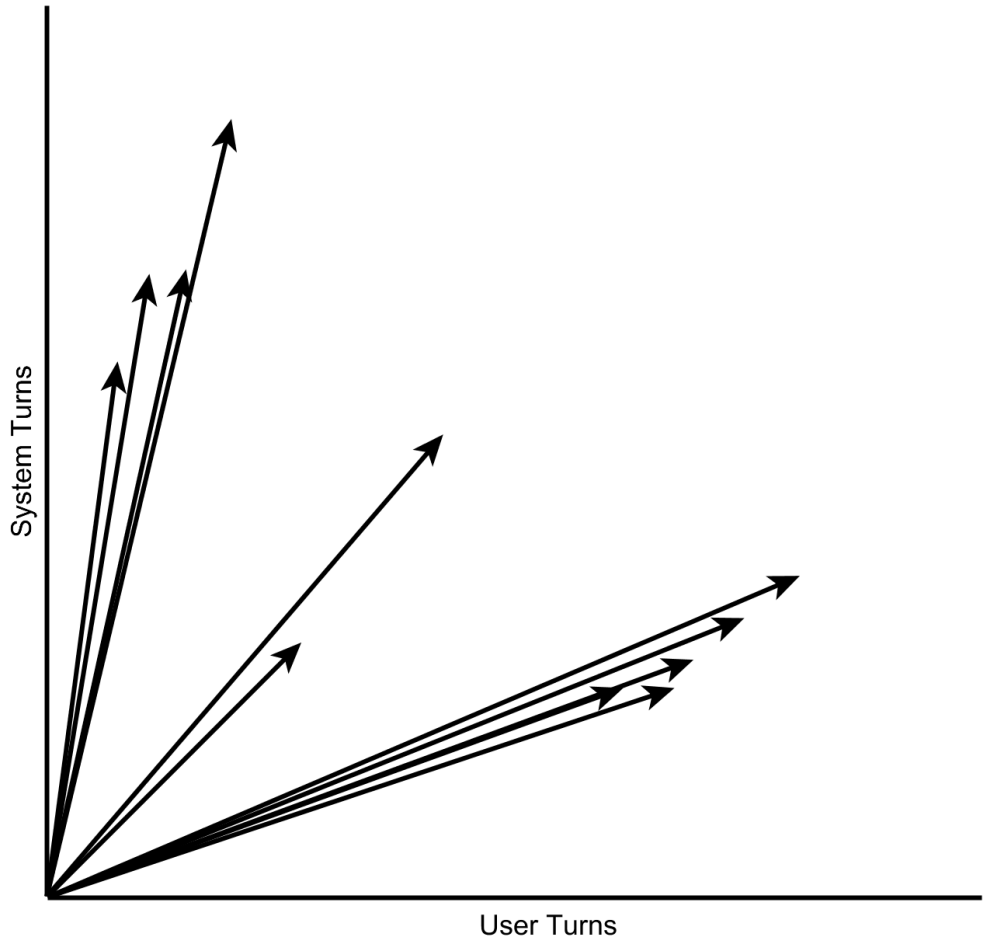
1. Extract a feature vector for each dialog
2. Create dialog clusters
3. Build linear regression models for the clusters
4. Give an unseen dialog a feature vector
5. Assign the dialog into a cluster
6. Use the cluster specific linear regression model to predict user satisfaction

Colaborative filtering

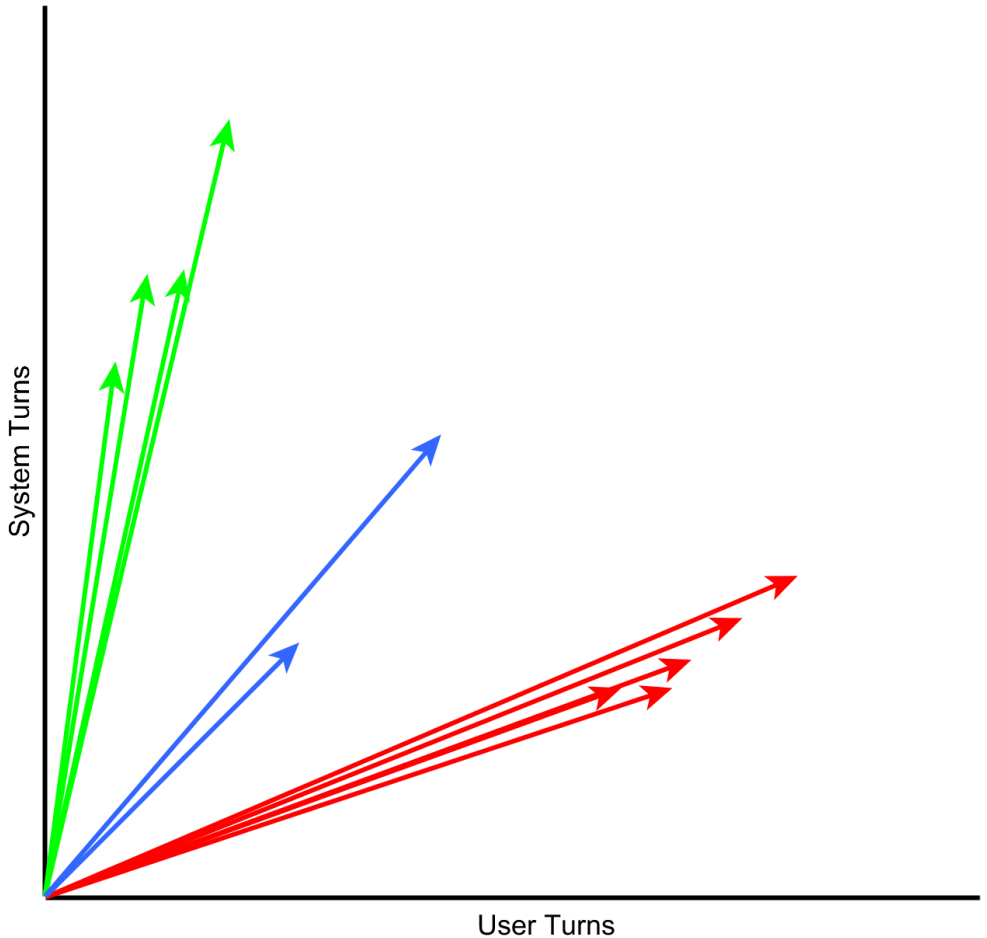
Table 2. *Features automatically extracted from log files.*

Feature	Definition
#System Turns	Overall number of system turns
#User Turns	Overall number of user turns
WPUT	Average number of words per user turn
AveUserSpeakRate	Average speaking rate of user's
AveRecogScore	Average recognition score
#Barge In	Overall number of user's barge in attempts
#Help Requests	Overall number of user's help requests
#User Questions	Overall number of user's questions
#System Questions	Overall number of system's questions
#DTMF	Overall number of touch tone uses

Colaborative filtering



Colaborative filtering



Thank you for your attention!

References

Paek, T. (2001, July) Empirical methods for evaluating dialog systems. Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9 (p. 2). Association for Computational Linguistics.

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997, July). PARADISE: A Framework for evaluating spoken dialogue agents. In Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 271-280). Association for Computational Linguistics.

Z. Yang, B. Li, Y. Zhu, I. King, G. Levow and H. Meng, Collaborative filtering model for user satisfaction prediction in Spoken Dialog System evaluation Spoken Language Technology Workshop (SLT), 2010 IEEE, Berkeley, CA, 2010, pp. 472-477.

Jokinen, K., & McTear, M. (2010). Spoken dialog systems. Morgan & Claypool, San Rafael.