Evaluating Task Success in a Dialogue System for Indoor Navigation

Nina Dethlefs, Heriberto Cuayáhuitl, Kai-Florian Richter,
Elena Andonova, John Bateman
University of Bremen, Germany
dethlefs@uni-bremen.de

Abstract

In this paper we address the assessment of dialogue systems for indoor wayfinding. Based on the PARADISE evaluation framework we propose and evaluate several task success metrics for such a purpose. According to correlation and multiple linear regression analyses, we found that task success metrics that penalise difficulty in wayfinding are more informative of system performance than a success/failure binary task success metric.

1 Introduction

Wayfinding in (partially) unknown environments poses a considerable challenge for humans. Our work addresses indoor navigation within complex buildings that present significant navigational challenges to new and infrequent visitors. This application scenario is of increasing relevance nowadays, as building complexes become larger and greater attention is paid to making them accessible for a broader range of users. Despite this, automatic systems generating natural language-based in-advance route descriptions have received little attention to date. Previous work consists of either attempts to involve primarily visual support or the quality and effects of the language component are taken for granted (Kray et al., 2005; Callaway, 2007; Kruijff et al., 2007). In particular, there appears to be a lack of agreed-on evaluation metrics for assessing the performance of dialogue systems in the wayfinding domain. This paper addresses this lack and suggests to employ metrics that are sensitive to task difficulty for evaluation of such systems. We present a study that confirms the usefulness of our proposed metrics.

2 System Architecture

We base our study on data collected from a dialogue system that assists users in indoor wayfinding using text-based natural language input and output. The system architecture consists of four different modules for the tasks of Natural Language Understanding, dialogue management, route instruction generation and Natural Language Generation. The following is a sample dialogue translated from German (S=System, U=User):

```
S: Do you need a route description on this
    level? Just write your question into
    the text field.
U: Where can I find room number 3180?
S: I have found the following rooms: b3180,
   a3180. Which are you looking for?
U: b3180
S: Please turn around and go straight until
   the next hallway on the left-hand side.
   Turn left and go until the door b3180
   on the right-hand side.
   Do you have further questions?
```

Since the task success metrics proposed in this paper aim to be independent of specific architectural decisions, we refer the reader to Cuayáhuitl et al. (2010) for details on our indoor navigation dialogue system.

3 Experimental Setting

3.1 Evaluation methodology

Evaluation of the system was performed using objective and subjective metrics mostly derived from the PARADISE framework (Walker et al., 2000). We used the following quantitative metrics. First, the group of *dialogue efficiency* metrics includes 'system turns', 'user turns', and 'elapsed time' (in seconds). The latter includes the time used by both conversants, from the first user utterance until the last system utterance. Second, the group of *dialogue quality* metrics consists of percentages of parsed sentences, sentences with spotted keywords, and unparsed sentences. Third, the group of *task success* metrics includes the well known success/failure Binary Task Success (BTS) defined as

$$\text{BTS} = \begin{cases} 1 & \text{for} \quad \text{Finding the Target Location (FTL),} \\ & \qquad\;\; \text{with or without problems} \\ 0 & \qquad\;\; \text{otherwise.} \end{cases}$$

Because this metric does not penalise difficulty in wayfinding, we propose and evaluate the following metrics — referred to as Graded Task Success (GTS) — that penalise with different values:

$$\text{GTS}^a = \begin{cases} 1 & \text{for} \quad \text{FTL without problems} \\ 0 & \qquad\;\; \text{otherwise,} \end{cases}$$

$$\text{GTS}^b = \begin{cases} 1 & \text{for} \quad \text{FTL with none or small problems} \\ 0 & \qquad\;\; \text{otherwise,} \end{cases}$$

$$\text{GTS}^c = \begin{cases} 1 & \text{for} \quad \text{FTL without problems} \\ 1/2 & \text{for} \quad \text{FTL with small problems} \\ 0 & \qquad\;\; \text{otherwise,} \end{cases}$$

$$\text{GTS}^d = \begin{cases} 1 & \text{for} \quad \text{FTL without problems} \\ 2/3 & \text{for} \quad \text{FTL with small problems} \\ 1/3 & \text{for} \quad \text{FTL with severe problems} \\ 0 & \qquad\;\; \text{otherwise.} \end{cases}$$

We coded difficulty in wayfinding, using the categories ‘no problems’, ‘small problems’ and ‘severe problems’ as follows. The value of 1 was given when the user finds the target location without hesitation, the value with ‘small problems’ was given when the user finds the location with slight confusion(s), and the value with ‘severe problems’ was given when the user gets lost but eventually finds the target location. The motivation behind using task success metrics that penalise differently the difficulty in wayfinding was to discover a metric that correlates highly with user satisfaction. Such a metric aims to be more informative for assessing task success performance than the traditional binary task success metrics. We tried four different graded metrics, GTS^a - GTS^d, in order to find the metric that best predicted user satisfaction. For the qualitative evaluation we used the subjective metrics described in (Walker et al., 2000).

3.2 Evaluation setup

Twenty-six native speakers of German participated in our study with an average age of 22.5 and a gender distribution of 16 female (62%) and 10 male (38%). Each subject received six dialogue tasks, corresponding to locations to find, which resulted in a total of 156 dialogues. Dialogues consisted of differing numbers of High-Level instructions (HLIs). High-Level Instructions (HLIs) encapsulate a set of low-level instructions (e.g., ‘go straight’, ‘turn left’, ‘turn around’) and are based on major direction changes. Two dialogue tasks used 2 High-Level Instructions (HLIs) such as those shown in the dialogue on page 1. Two other tasks used 3 HLIs, and two used 4 HLIs. The tasks were executed pseudorandomly (from a uniform distribution), so that the order of task execution would not impact on the user ratings. The participants were asked to request a route from the system using natural language, optionally take notes, and then follow the system instructions closely trying to find the locations. They were not allowed to ask anybody for help. Participants could give up when they were unable to find the target location by telling that to the assistant that followed them. It was the task of this assistant as well to judge and take note of the difficulties that subjects encountered in their wayfinding task as described in the previous section. At the end of each dialogue, participants were asked to fill in a questionnaire for obtaining qualitative results using a 5-point Likert scale, where 5 represents the highest score.

4 Experimental Results

Table 1 summarises our results for the quantitative and qualitative metrics. It can be observed from the dialogue efficiency metrics (first group) that user-machine interactions involved short dialogues in terms of turns and interaction time. Once users received instructions from the system, they tended not to ask further. With regard to dialogue quality (second group), we noted that our grammars need to be extended in coverage and that the keyword spotter proved vital in the dialogues. The analysis of task success measures (third group) revealed very high binary task success, and lower scores for the other task success metrics.

4.1 Correlation analysis

In a correlation analysis between task success measures and user satisfaction we obtained the results displayed in Table 2. This can be interpreted as follows: while all metrics correlate moderately with overall user satisfaction, the metrics taking task difficulty into account correlate higher. A more detailed analysis of the corre-

Table 1: Mean values of our evaluation metrics for our wayfinding system based on 156 dialogues, organised in four groups: dialogue efficiency, dialogue quality, task success and user satisfaction.

Measure	Score
System Turns	2.30 ± 0.3
User Turns	1.52 ± 0.5
System Words per Turn	41.30 ± 4.0
User Words per Turn	4.79 ± 2.1
Interaction Time (secs.)	22.14 ± 18.4
Session Duration (secs.)	2014.62 ± 393.2
Parsed Sentences (%)	16.7 ± 16.0
Spotted Keywords (%)	79.9 ± 17.0
Unparsed Sentences (%)	3.4 ± 0.5
Binary Task Success (%)	94.9 ± 8.3
Graded Task Success^a (%)	71.4 ± 15.0
Graded Task Success^b (%)	87.8 ± 15.0
Graded Task Success^c (%)	81.4 ± 13.3
Graded Task Success^d (%)	87.6 ± 8.3
(Q1) Easy to Understand	4.46 ± 0.8
(Q2) System Understood	4.65 ± 0.8
(Q3) Task Easy	4.29 ± 0.9
(Q4) Interaction Pace	4.63 ± 0.5
(Q5) What to Say	4.66 ± 0.7
(Q6) System Response	4.56 ± 0.6
(Q7) Expected Behaviour	4.45 ± 0.8
(Q8) Future Use	4.31 ± 0.9
Overall User Satisfaction (%)	90.0 ± 7.3

lation between task success metrics and individual user satisfaction metrics revealed the following. First, the binary task success showed lower correlations than the other metrics in the subjective metric 'easy to understand' (Q1). Second, while there is no correlation between the subjective metric 'future use' (Q8) and binary task success, the other metrics reveal a moderate correlation. Third, while binary task success shows a moderate correlation for 'task easy' (Q3), the other metrics show a high correlation. Therefore, we can conclude that the task success metrics that penalise difficulty in wayfinding are more informative of user-system interaction performance for indoor wayfinding than the BTS metric. Furthermore, there was no correlation between the number of high-level instructions and overall user satisfaction, i.e. user satisfaction was independent of instruction length (our system performed equally well for short and long routes).

Table 2: Correlation coefficients between task success and user satisfaction measures (significant at $p < 0.05$).

Measure	BTS	GTS^a	GTS^b	GTS^c	GTS^d
Q1	.47	.44	.54	.49	.54
Q2	.20	.17	.19	.19	.20
Q3	.53	.67	.71	.71	.76
Q4	.21	.26	.24	.24	.28
Q5	.20	n.s.	.17	.18	.18
Q6	n.s.	n.s.	n.s.	n.s.	n.s.
Q7	.31	.35	.44	.40	.44
Q8	n.s.	.39	.32	.40	.39
Overall	.43	.52	.55	.55	.60

Note: n.s. - not significant.

4.2 Multiple linear regression analysis

In order to identify the relative contribution that different factors have on the variance found in user satisfaction scores, we performed a standard multiple linear regression analysis on our data. According to the PARADISE framework (Walker et al., 1997), performance can be modeled as a weighted function of task-success measure and dialogue-based cost measures. The latter represent the measures summarised under dialogue efficiency and dialogue quality above. We normalised all task success and cost values to account for the fact that they can be measured on different scales (seconds, percentages, sum, etc.), according to $\mathcal{N}(x) = \frac{x-\bar{x}}{\sigma_x}$, where σ_x corresponds to the standard deviation of x. Then we performed several regression analyses involving these data.

Results revealed that the metrics 'user turns' and 'task success' (for GTS^a, GTS^c and GTS^d) were the only predictors of user satisfaction at $p < 0.05$. The other task success measures were not significant (with BTS at $p = 0.39$ and GTS^b at $p = 0.17$). These results confirm our claim that task success metrics that consider difficulty in wayfinding (specifically GTS^a, GTS^c and GTS^d) are more informative with respect to user satisfaction in the wayfinding domain than a binary success/failure metric. Subjects seem to be sensible to problems they encounter in their wayfinding tasks, which are expressed in their ratings of the system.

4.3 Estimation of a performance function

We use the following equation to obtain a performance function (Walker et al., 1997):

$$\text{Performance} = (\alpha * \mathcal{N}(k)) - \sum_{i=1}^{n} \omega_i * \mathcal{N}(c_i),$$

where, α is a weight on the task success metric k (to be replaced by any of our proposed metrics), and ω_i is a weight on the cost functions c_i. $\mathcal{N}$ represents the normalised value of c_i. Based on the results of our first regression analysis, we ran a second analysis using those variables that were significant predictors in the first regression, i.e. the number of user turns and task success metrics GTS^a, GTS^c and GTS^d. We analysed the correlation between these variables, which resulted in weak negative correlations. We obtained the following performance function for task success metrics GTS^c and GTS^d (because those two accounted for most of the variance in user satisfaction), where UT refers to 'User turns':

$$\text{Performance} = 0.38\mathcal{N}(GTS^{c,d}) - 0.87\mathcal{N}(UT),$$

suggesting that the more successful and efficient the interaction, the better. These results show that GTS^c, GTS^d and UT are significant at $p < 0.01$, and the combination of UT and each of GTS^c and GTS^d account for 62% of variance in user satisfaction. This performance function can be used in future evaluations of the system.

5 Discussion

The idea of taking different degrees of task difficulty into consideration in evaluation is not entirely new (Tullis and Albert, 2008). However, to the best of our knowledge, there have been no previous studies that demonstrated that these metrics do indeed show a higher correlation with user satisfaction scores than the BTS metric, which is typically used to assess task success. This finding was supported by an evaluation in a real environment using an end-to-end dialogue system, and was based on PARADISE, a generic framework for the evaluation of (spoken) dialogue systems. The proposed metrics can therefore be regarded as a useful and important step contributing to the understanding of the performance of situated dialogue systems. Further, our proposed metrics address the lack of standardised evaluation metrics in the wayfinding domain in particular. We presented a concrete performance function that can help future system development in the domain by allowing the estimation of relative contributions of different task success metrics and cost function towards overall user satisfaction.

6 Conclusion

In this paper we addressed the assessment of dialogue systems for indoor navigation using the PARADISE framework and different task success metrics. We found that task success metrics that take difficulty in wayfinding into account correlate higher with overall user satisfaction than a binary task success metric. In addition, a more detailed correlation analysis for subjective metrics of user satisfaction confirmed that our proposed metrics are more informative of system performance for indoor wayfinding than the binary success/failure metric. This result was confirmed by a multiple linear regression analysis that tested for the relative contribution to variance in user satisfaction of different task success metrics and cost measures. Future work can apply these metrics to dialogue systems with different input and output modalities.

Acknowledgements

Part of this work was supported by the DFG Transregional Collaborative Research Center SFB/TR8 "Spatial Cognition".

References

Charles Callaway. 2007. Non-localized, interactive multimodal direction giving. In I. van der Sluis, M. Theune, E. Reiter, and E. Krahmer, editors, *Proceedings of the Workshop on Multimodal Output Generation MOG 2007*, pages 41–50. Centre for Telematics and Information Technology (CTIT), University of Twente.

Heriberto Cuayáhuitl, Nina Dethlefs, Kai-Florian Richter, Thora Tenbrink, and John Bateman. 2010. A dialogue system for indoor wayfinding using text-based natural language. In *International Journal of Computational Linguistics and Applications, ISSN 0976-0962*.

Christian Kray, G. Kortuem, and A. Krüger. 2005. Adaptive navigation support with public displays. In Robert St. Amant, John Riedl, and Anthony Jameson, editors, *Proceedings of IUI 2005. ACM Press, New York*, pages 326–328.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138, March. Special Issue on Human-Robot Interaction.

Tom Tullis and Bill Albert. 2008. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *ACL*, pages 271–280.

M. Walker, C. Kamm, and D. Litman. 2000. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3):363–377.