Empirical Evaluation for Hindi text-to-scene generation system

Priyanka Jain[*], Ram Bhavsar[**], B. V. Pawar[**], Hemant Darbari[*]
[*]Center for Development of Advanced Computing, Pune, India
[**]North Maharashtra University, Jalgaon, India

Abstract : *We have developed Preksha: a Hindi text visualizer. Preksha* is an automatic text visualization [ATV] system. It *generates a 3D scene by understanding a given Hindi text.* This is the process of natural language understanding and *conversion of this textual information into 3D scene generation. In this paper, we discuss the challenges for evaluation of ATV system. The area of text visualization is comparative new in discipline. It does not have much proven work toward the finalizing standards and strategies for Evaluation process. We present an evaluation plan for Preksha and* initial results of Preksha-Evaluation.

***IndexTerms* - Artificial Intelligence, text-to-conversion, Natural Language Visualization, Human Computer Interface, Natural language processing, Hindi Language.**

I. INTRODUCTION

The purpose of this research is to understand and prepare an initial plan for ‘Preksha’. Preksha is an Automatic Text Visualization (ATV) system, which is the only reported work for language Hindi. Preksha comprehends the information present in a Hindi text to transform into a 3D scene form. This requires natural language processing, knowledge processing and scene generation processing. The implementation of Preksha is done based on previous reported works in [7], [8], [9], [10], and [11]. This paper presents the evaluation strategy for Preksha. Section 2 presents the challenges in evaluation processes of ATV systems. Preksha Evaluation plan is discussed in Section 3.Section 4 is discusses the experiments and results. Section 5 is conclusion.

II. ATV EVALUATION CHALLENGES

Evaluation design methods are cognitive walkthrough, heuristic evaluation and review based evaluation [1]. Review based evaluation is beneficial as these are quick, reach large user group and can be analyzed more rigorously. Its disadvantages are that these are less flexible and need careful design. This approach involves users with experimental methods, observational methods and query methods. A questionnaire [6] is a measurement tool designed to assess a user's subjective contentment with an interface. This list of questions provided to users for quantitative response. Types of questions in questionnaires are general, open-ended, scalar, multi-choice, or ranked.

In absence of standard evaluation methodologies for text visualization, we define our plan for ATV evaluation considering the complexity of system. Evaluation of an ATV system is an equally difficult task as building them; this fact has been underlined in the operational challenges. There are various factors, which governs the evaluation of multi-dimensional ATV system. Unlike text-to-text machine translation system, we cannot count matched and unmatched words for Word Error Rate (WER) and MWER (MWER) [12] in text-to-scene translation system. The output of ATV system is 3D scene. Building references by manual scene creation and comparing it with the machine output for automatic evaluation is also not feasible. The reason behind this is that the visualization has almost unlimited possibility of reference scene generation. Both in case of mental and machine output, comparing semantic of scene is almost impossible in today’s technical world. In this case, the justification for evaluating the machine-generated output with one-of-the-many possibility of references is challenging.

The measurement of the success is very subjective in case of automatic visualization. Different human interpretations of input text and corresponding visualization of virtual world is another challenges. The output generated is visual in nature and is not certain to choose features in a generated scene for measurement as a quantitative evaluation. Most related research performs subjective evaluations using visual examples [4], [2], [3] and [5]. In these cases, a small set of example images produced from the original text is provided, leaving evaluation to the discretion of the reader.

III. PREKSHA EVALUATION

Evaluation techniques in general fall into two categories: subjective techniques and objective techniques. Subjective techniques require the participation of human subjects but objective techniques does not require that. Both subjective and objective techniques can be used to provide either global or local evaluation of language visualization. Since working examples of other systems are limited, the evaluation does not include relative evaluation, i.e. comparison with other systems. In this evaluation plan, we investigate the following questions:

1) Can our implementation of underlying algorithms be used to locate solutions to non-quantified benchmarks?
2) Are consistent high-level scene descriptions derived from annotated text?
3) Are virtual environments populated consistently using high-level scene descriptions?
4) Is the automatically populated 3D environments representative of the corresponding text?

This question cannot be answered using quantified methods, due to the subjectivity of language and visual interpretation. However, we provide examples of automatically created environments, and provide a subjective evaluation of each. Evaluation strategy is planned with different social group and different approaches like -technical and human evaluation.

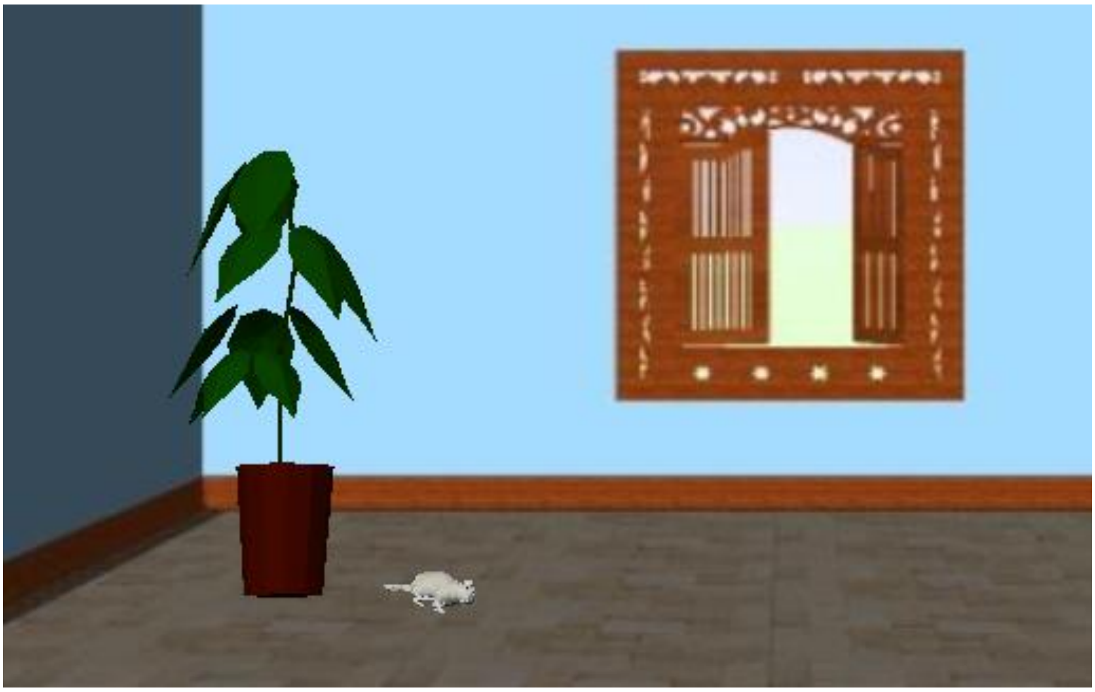
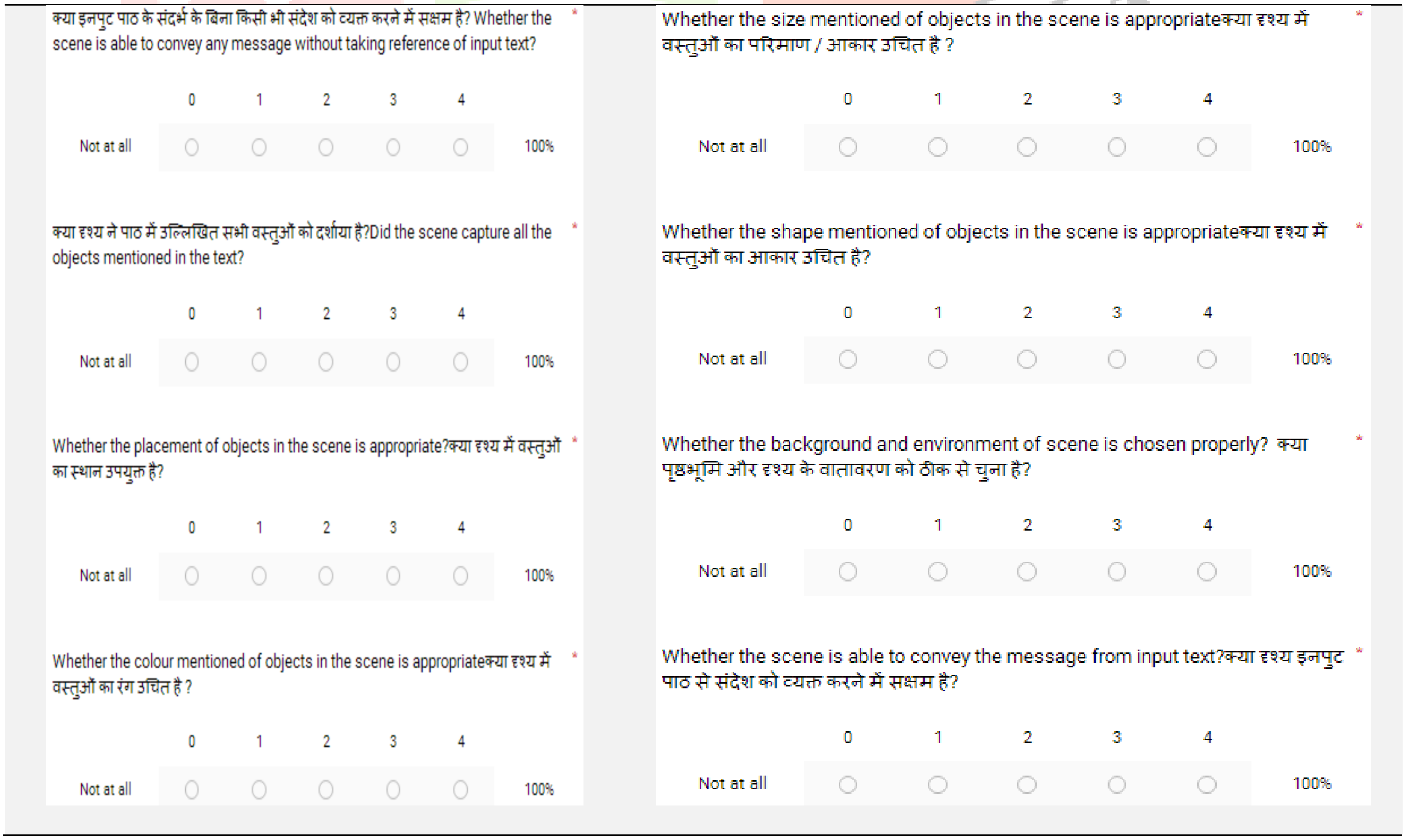


Figure 1. Scene generated from Preksha for sample sentence कमरे में गमले के पास चूहे हैं [There are mice near the flowerpot in room]

IV. Subjective Evaluation

Testing Preksha on open test corpora in the limited scope of research is very difficult, as it demands with heavy linguistic & lexical preparation. Hence, it is decided to test Preksha with test suite sentences designed within the scope with couple of simple verbs, which are easy for visualization. The scope is also limited to fiction-based simple sentences of indoor text describing few objects. We evaluate Preksha in a questionnaire experiment. We evaluate the output of our system by asking people to judge the matching of generated scenes with given input descriptions. Evaluator groups are formed pertaining to various age groups and social group whose may and may not be equipped with computer environment. All members participated on a voluntary basis. The age range of the participants was 10-50 year. An initial introduction to Preksha research and a skill building session is been taught.

A questioner is prepared and circulated as a Google doc form. This questioner is prepared using Ten (10) stories and their corresponding scene generated using Preksha system. A basic introduction of research and evaluation objective was presented before starting of the survey. After receiving basic information of evaluator, a questionnaire was presented to evaluator associated to scene understanding. Evaluators fill the form with their basic information then provide their answers of question asked. Figure 2 presents a sample questioner used for Preksha evaluation.



क्या इनपुट पाठ के संदर्भ के बिना किसी भी संदेश को व्यक्त करने में सक्षम है? Whether the scene is able to convey any message without taking reference of input text? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

क्या दृश्य ने पाठ में उल्लिखित सभी वस्तुओं को दर्शाया है?Did the scene capture all the objects mentioned in the text? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Whether the placement of objects in the scene is appropriate?क्या दृश्य में वस्तुओं का स्थान उपयुक्त है? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Whether the colour mentioned of objects in the scene is appropriateक्या दृश्य में वस्तुओं का रंग उचित है ? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Whether the size mentioned of objects in the scene is appropriateक्या दृश्य में वस्तुओं का परिमाण / आकार उचित है ? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Whether the shape mentioned of objects in the scene is appropriateक्या दृश्य में वस्तुओं का आकार उचित है? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Whether the background and environment of scene is chosen properly? क्या पृष्ठभूमि और दृश्य के वातावरण को ठीक से चुना है? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Whether the scene is able to convey the message from input text?क्या दृश्य इनपुट पाठ से संदेश को व्यक्त करने में सक्षम है? *

	0	1	2	3	4	
Not at all	○	○	○	○	○	100%

Figure 2. Sample questioner

Evaluators were asked to rank the eight answers of 10 stories on the scale of 0 to 4. Interpretation of these scales is explained in table 1 as "4-point Intelligibility/Accuracy Human Evaluation Tests Score Sheet." Table 2 presents the scope of accuracy or fidelity test which were not carried out in case of this research work.

Table 1. A 4-point Intelligibility Test scale Score Sheet

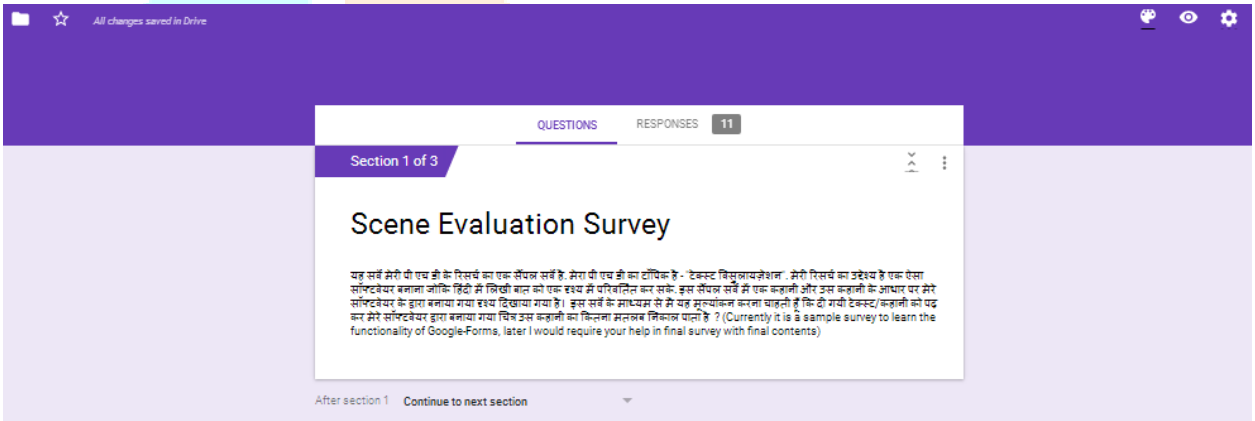
Scale	Meaning	Interpretation
3	Very intelligible	Grammatically correct and quite clear
2	Fairly intelligible	Generally clear with few inaccuracies but information conveyed is understandable
1	Barely intelligible	Central idea is clear only after considerable study, full of errors and poor word choice
0	Unintelligible	Nothing can be understood after any amount of efforts. Completely weird.

Table 2. A 4-point Accuracy/Fidelity Test Score sheet

Scale	Meaning
3	Completely faithful
2	Fairly intelligible, more than 50% of the original information is passed on
1	Barely intelligible, less than 50% of the original information is passed on
0	Completely unfaithful, doesn't make sense at all

Each story is evaluated by 11 numbers of evaluators, and taking average of these ranking estimates the evaluation result of that particular story. Figure 3, 4, 5, 6 and 7 presents the evaluation process of Preksha.

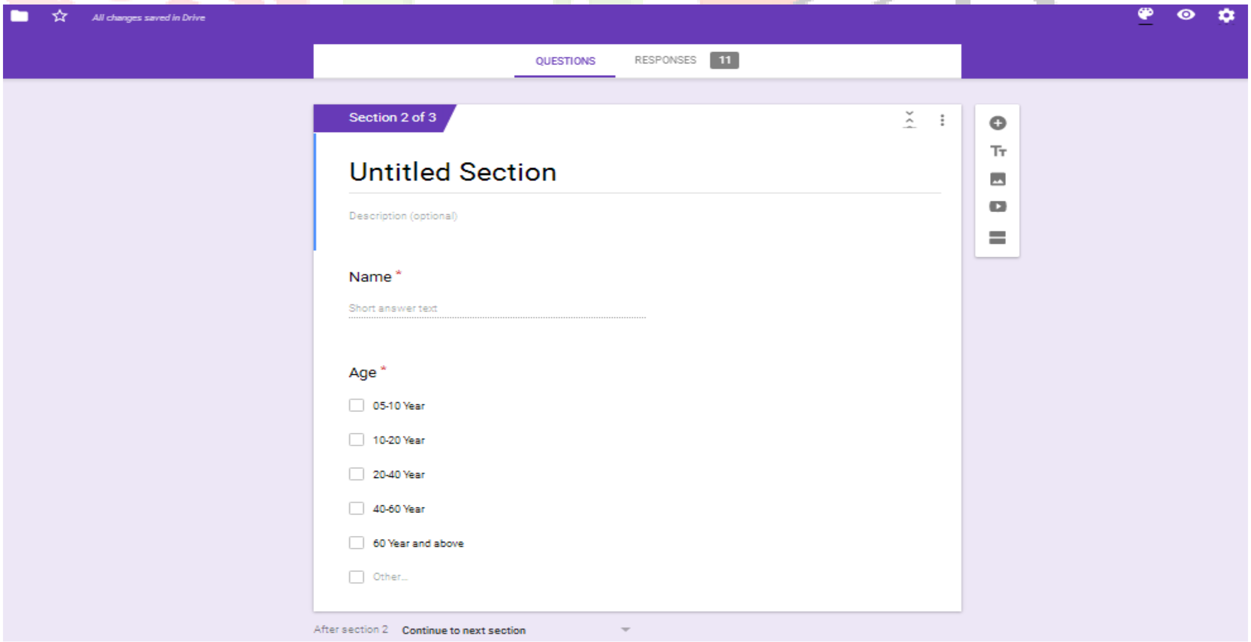
In this way, we calculate the evaluation output of all 10 stories. The grand average of all these 10 stories rank provides the final evaluation result of Preksha System. Table 3 shows the final results grade for Preksha evaluation.



Scene Evaluation Survey

यह सर्वे मेरी पी एच डी के रिसर्च का एक सैंपल सर्वे है. मेरा पी एच डी का टॉपिक है - 'टेक्स्ट विसुलायज़ेशन'. मेरी रिसर्च का उद्देश्य है एक ऐसा सॉफ्टवेयर बनाना जोकि हिंदी में लिखी बात को एक दृश्य में परिवर्तित कर सके. इस सैंपल सर्वे में एक कहानी और उस कहानी के आधार पर मेरे सॉफ्टवेयर के द्वारा बनाया गया दृश्य दिखाया गया है। इस सर्वे के माध्यम से मैं यह मूल्यांकन करना चाहती हूँ कि दी गयी टेक्स्ट/कहानी को पढ़ कर मेरे सॉफ्टवेयर द्वारा बनाया गया चित्र उस कहानी का कितना मतलब निकाल पाता है ? (Currently it is a sample survey to learn the functionality of Google-Forms, later I would require your help in final survey with final contents)

After section 1 Continue to next section



Untitled Section

Description (optional)

Name *

Short answer text

Age *

□ 05-10 Year

□ 10-20 Year

□ 20-40 Year

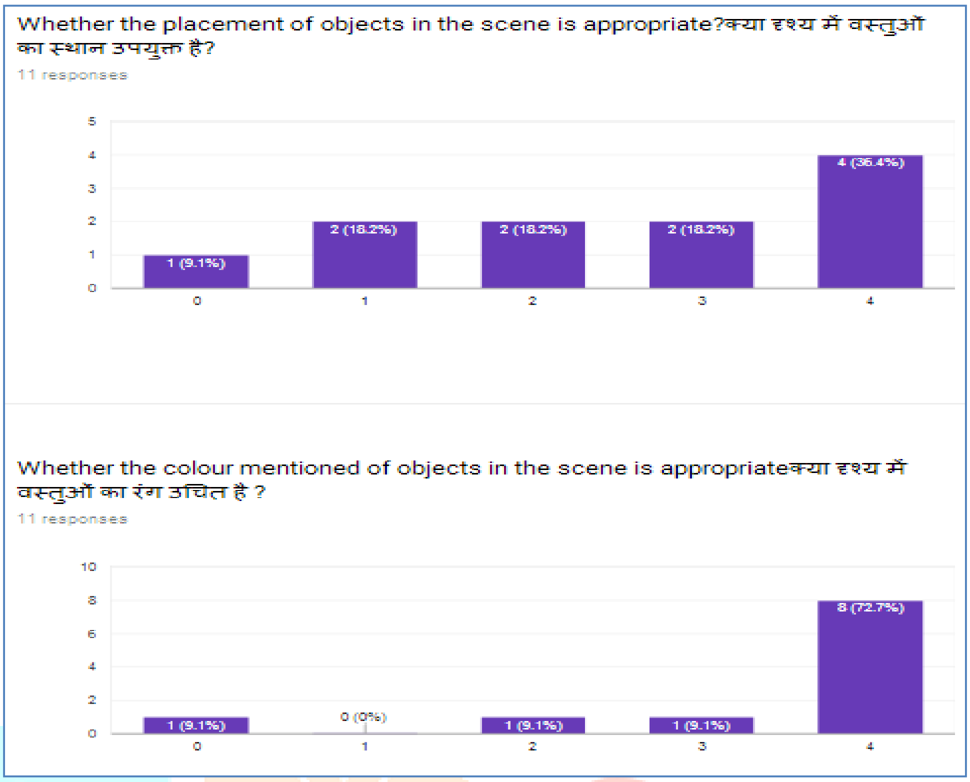
□ 40-60 Year

□ 60 Year and above

□ Other...

After section 2 Continue to next section

Figure 3. Introductory information on google-doc for Preksha Evaluation



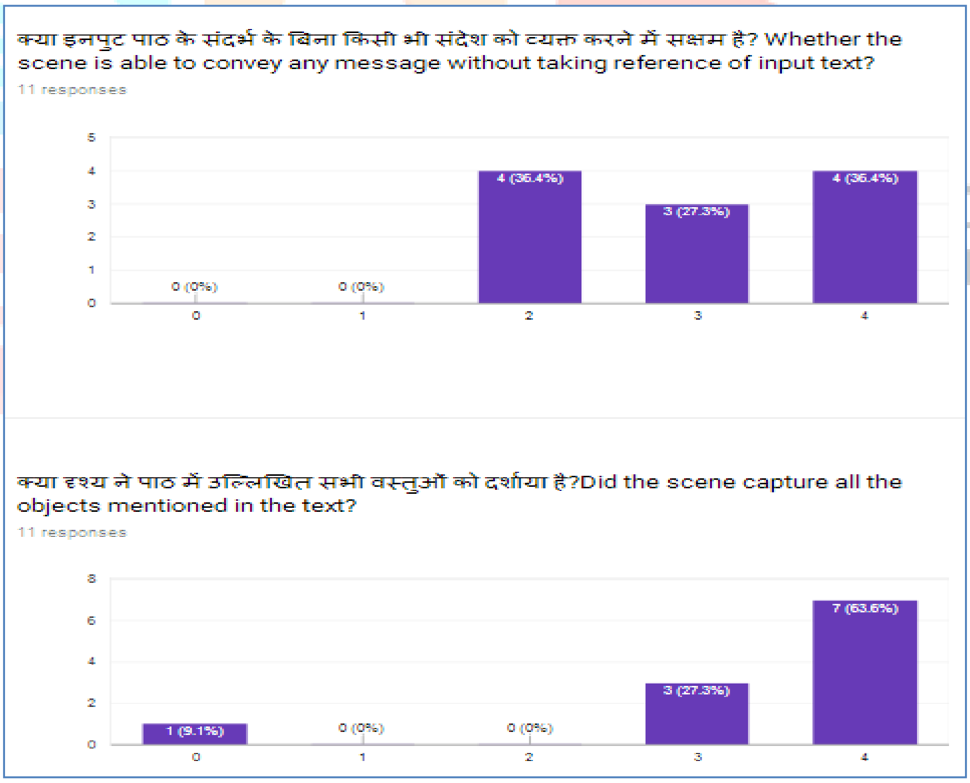
Whether the placement of objects in the scene is appropriate?क्या दृश्य में वस्तुओं का स्थान उपयुक्त है?

11 responses

Whether the colour mentioned of objects in the scene is appropriateक्या दृश्य में वस्तुओं का रंग उचित है ?

11 responses

Figure 4. Preksha Evaluation graph for first 2 questions



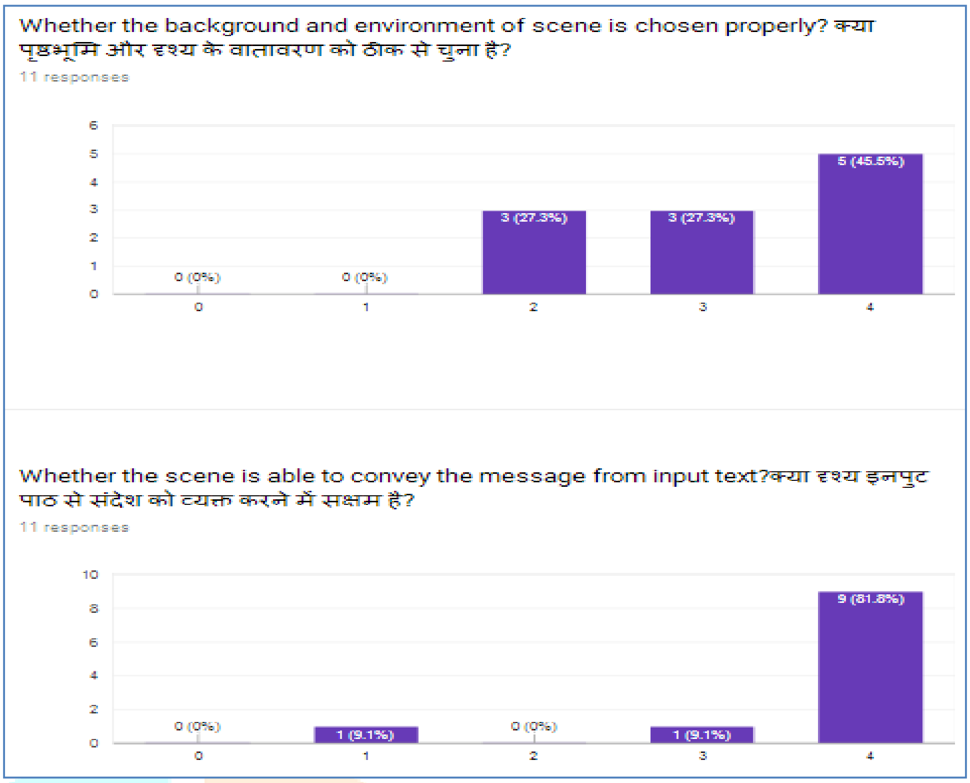
क्या इनपुट पाठ के संदर्भ के बिना किसी भी संदेश को व्यक्त करने में सक्षम है? Whether the scene is able to convey any message without taking reference of input text?

11 responses

क्या दृश्य ने पाठ में उल्लिखित सभी वस्तुओं को दर्शाया है?Did the scene capture all the objects mentioned in the text?

11 responses

Figure 5. Preksha Evaluation graph for next 2 questions



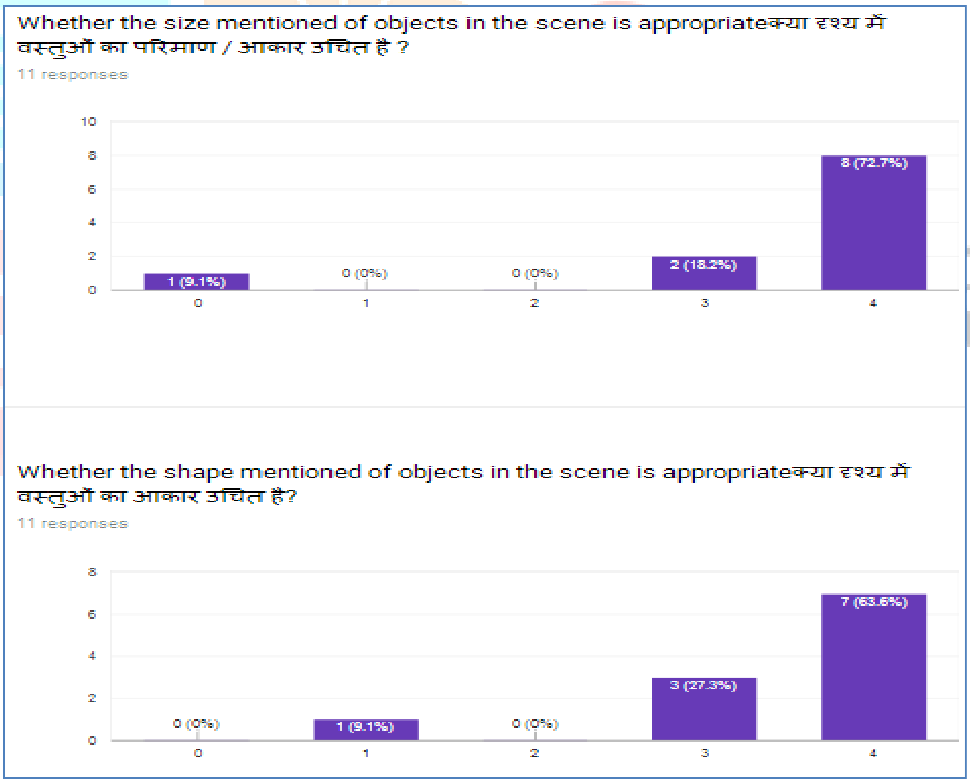
Whether the background and environment of scene is chosen properly? क्या पृष्ठभूमि और दृश्य के वातावरण को ठीक से चुना है?

11 responses

Whether the scene is able to convey the message from input text?क्या दृश्य इनपुट पाठ से संदेश को व्यक्त करने में सक्षम है?

11 responses

Figure 5. Preksha Evaluation graph for next 2 questions



Whether the size mentioned of objects in the scene is appropriateक्या दृश्य में वस्तुओं का परिमाण / आकार उचित है ?

11 responses

Whether the shape mentioned of objects in the scene is appropriateक्या दृश्य में वस्तुओं का आकार उचित है?

11 responses

Figure 6. Preksha Evaluation graph for last 2 questions

V. Result Analysis

We have set forth a subjective evaluation methodology which tests the intelligibility of the Preksha system. This is a initial evaluation where we find the appropriate approach for calculating evaluation results based on different parameters.

Table 3. Evaluation Results from Preksha

Date/Time	Name	Age (Year)	Que 1	Que 2	Que 3	Que 4	Que 5	Que 6	Que 7	Que 8	Total	Average
2017/09/16 1:24:22	User 1	10-20	2	4	1	3	4	4	4	4	26	3.25
2017/09/16 1:47:05	User 2	40-60	3	3	3	4	4	3	3	4	27	3.375
2017/09/16 2:02:20	User 3	10-20	2	3	4	4	4	4	4	4	29	3.625

2017/09/16 2:24:59	User 4	10-20	2	4	2	4	3	4	2	4	25	3.125
2017/09/16 2:25:21	User 5	10-20	2	4	2	4	3	4	2	4	25	3.125
2017/09/16 3:16:36	User 6	20-40	4	0	0	0	0	1	2	1	8	1
2017/09/16 3:49:42	User 7	10-20	4	4	4	4	4	4	4	4	32	4
2017/09/16 4:19:50	User 8	20-40	4	4	4	4	4	4	4	3	31	3.875
2017/09/16 4:21:16	User 9	20-40	4	4	4	4	4	4	4	4	32	4
2017/09/16 4:51:14	User 10	20-40	3	4	1	4	4	3	3	4	26	3.25
2017/09/16 7:24:11	User 11	40-60	3	3	3	2	4	3	3	4	25	3.125
		Total	33	37	28	37	38	38	35	40	286	**3.25**
		Average	3	3.36	2.54	3.36	3.45	3.4	3.18	3.6	**3.25**	

We calculate the evaluation output of all 10 stories. The grand average of all these 10 stories rank provides the final evaluation result of Preksha System. Table 3 shows the final results grade for Preksha evaluation. By mentioned evaluation strategy, Preksha scores 3.25 / 4 in scale.

VI. CONCLUSION

The area of text visualization is comparative new in discipline. It does not have much proven work toward the finalizing standards and strategies for Evaluation process. We have offered a subjective evaluation process which is more suitable for human computer interaction systems where automatic evaluation is not set final. This research provides a wider future scope for researchers with new methodologies to put forward. After a short discussion on evaluation challenges and design methods, this paper describes the evaluation process, and evaluation results.

REFERENCES

[1] Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces, Proc. ACM CHI'90 Conf. (Seattle, WA, 1-5 April), 249-256.

[2] Lu, R. and Zhang, S.: Automatic Generation of Computer Animation – Using AI for Movie Animation. LNCS 2160, NY: Spinger-Verlag, pp. 1-374, (2002).

[3] Zeng, X., Mehdi, Q. H., and Gough, N. E.: Shape of the story: Story visualization techniques. In IV '03: Proceedings of the 7th International Conference on Information Visualization (London, United Kingdom), pp. 144 (2003)

[4] Coyne, B., Sproat, R.: WordsEye: An automatic text-to-scene conversion system. SIGGRAPH, Computer Graphics Proceedings pp. 487-496, (2001).

[5] Joshi et al., 2004. Joshi, D., Wang, J.Z., and Li, J., 2004. The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. In Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval. New York, USA, pp. 1073–1078.

[6] John P. Chin, Virginia A. Diehl, Kent L. Norman, Development of a Tool Measuring User Satisfaction of the Human-Computer Interface, Department of Psychology, University of Maryland

[7] Jain, P., Bhavsar, R. P., Lele, A. Kumar, A., Pawar, B. P., Darbari, H., and Bhavsar, V. C.: "Knowledge acquisition for automatic text visualization" in National Conference on Advances in Computing (NCAC-2017). 2017

[8] Jain, P., Darbari, H., and Bhavsar, V, C.: 'Vishit: A Visualizer for Hindi Text'. pp. 886-890. Fourth International Conference on Communication Systems and Network Technologies (CSNT), IEEE Xplore. 2014.

[9] Jain, P., Darbari, H., and Bhavsar, V. C.: 'Text Visualization as an Aid to Language Learning Disability', pp. 88. ELELTECH 2013 National Conference on e-Learning and e-Learning Technologies, India. 2013.

[10] Jain, P., Darbari, H., and Bhavsar, V. C.: "Cognitive support by Language Visualization A case study with Hindi Language" in 2nd International Conference for Convergence in Technology (I2CT), IEEE Xplore. 2017.

[11] Jain, P., Darbari, H., and Bhavsar, V. C.: "Spatial Intelligence from Hindi Language Text for Scene Generation" in 2nd International Conference for Convergence in Technology (I2CT), IEEE Xplore. 2017.

[12] Klakow, Dietrich; Jochen Peters (September 2002). "Testing the correlation of word error rate and perplexity". Speech Communication. 38 (1-2): 19–28. doi:10.1016/S0167-6393(01)00041-3. ISSN 0167-6393. Retrieved 28 August 2013.