Production and evaluation of (multimodal) answers to medical questions

Charlotte van Hooijdonk[a], Emiel Krahmer[b], Alfons Maes[b], Mariët Theune[c], Wauter Bosma[a]
[a]VU University Amsterdam
[b]Tilburg University
[c]University of Twente
CMJ.VAN.HOOIJDONK@LET.VU.NL

This paper describes two experiments carried out to investigate the production and evaluation of multimodal answer presentations in the context of a medical question answering system. In a production experiment participants had to produce answers to different types of questions. The results show that about one in four produced answers using multiple media. In an evaluation experiment, users had to evaluate different types of multimodal answer presentations. Answers with an informative visual were evaluated as more informative and more attractive than answers with a mere illustrative visual.

Keywords: Multimodal information presentation, cognitive engineering, document design, visual media

Introduction

This paper investigates the production and evaluation of multimodal answer presentations in a medical question answering system (QA). Early QA research concentrated on textual answers to factoid questions (i.e., What is the capital of France? Paris). Currently, there is a growing interest in the generation of multimodal answers to more complex questions. This raises questions about what combinations of modalities are most appropriate given particular types of questions.

Multimodal information presentation has been studied in various research fields with various outcomes. For example, research in cognitive and educational psychology focused on how multimodal presentations affect the users' understanding, recall and processing efficiency of the presented material (e.g., Carney & Levin, 2002; Mayer, 2005; Tversky, Morrison, Bétrancourt, 2002). Guidelines resulting from this research often relate to specific types of information used in specific domains, for example cause and effect chains which explain how systems work (e.g., Mayer & Moreno, 2002) or procedural information (e.g., Michas and Berry, 2000). Research in language generation research has tried to classify and characterize modalities, information types, and the matches between them. For example, Bernsen (1994) proposed a taxonomy of generic unimodalities consisting of various features. Other scholars studied the so-called *media allocation problem* (i.e., How to determine which information to allocate to which medium) and tried to identify which factors play a role in media allocation (Arens, Hovy and Vossers, 1993).

In short, attempts have been made to generate optimal multimodal information presentations resulting in several modality guidelines, frameworks, and taxonomies. Still needed is information about people's modality preferences in producing and evaluating presentations. Therefore, we carried out two experiments following the approach of Heiser, Phan, Agrawala, Tversky and Hanrahan (2004), where people are asked to produce information presentations (e.g., assembly instructions), which are then rated by others.

Experiment I: Production

Participants and stimuli

111 students of Tilburg University participated for course credits (65 female, 19-33 years old). Participants were given one of four sets of eight medical questions for which the answers could be found on the Internet. Four were randomly chosen from one hundred medical questions formulated to

test the IMIX QA system. Of the remaining four questions, two were definition questions (e.g., "What does ADHD stand for?") and two were procedural questions (e.g., "How to apply a sling to the left arm?"). Participants had to give two answers per question, a brief and an extended answer, using whatever combinations of modalities they wanted. They were specifically asked to present the answers as they would prefer to find them in present day digital information environment. Questions and answers had to be presented in a fixed format in PowerPoint™ with areas for the question ('vraag') and the answer ('antwoord'). They were acquainted with inserting different types of objects in PowerPoint.

Coding system and procedure

Each answer was coded on the presence of visual media (photos, graphics, and animations) and on the function of these visual media in relation to the text, loosely based on Carney & Levin (2002), i.e., decorative, representational, or informative. In total 1775 answers were collected (111 participants × 8 questions × 2 answers, minus one missing answer). Six analysts independently coded the same set of 111 answers. Subsequently, every analyst independently coded a part of the total corpus (approximately 300 answers). Calculations of Cohen's κ showed that the analysts almost perfectly agreed in judging the occurrence of photos (κ = .81), graphics (κ = .83), and animations (κ = .92). An almost perfect agreement was reached in assigning the function of the visual media (κ = .83).

Results

Analysis of the complete corpus of coded answer presentations showed that almost one in four answers contained one or more visual media, of which graphics were most frequent (14,9%) and animations were least frequent (3,8%). The presence of photos was between these two (8,6%).

Table 1*: Percentages of visual media functions related to answer length (n = 442)*

	Brief answers (n = 101)	Extended answers (n = 341)
Decorative visuals (n = 70)	26.7	12.6
Representational visuals (n = 201)	20.8	52.8
Informative visuals (n = 171)	52.5	34.6

Table 1 shows that visual media occurred significantly more often with extended answers (χ^2 (1) = 173.89, p< .001). Moreover, the distribution of the functions of visual media differed significantly over answer length (χ^2 (2) = 33.79, p< .001). Informative visuals occurred more often in brief answers, whereas representational visuals occurred more often in extended answers.

Table 2*: Percentages of the functional types of visual media related to definition and procedural questions (n =271)*

	Definition questions (n = 91)	Procedural questions (n = 180)
Decorative (n = 27)	19.8	5.0
Representational (n = 129)	53.8	44.4
Informative (n = 115)	26.4	50.6

Table 2 shows that visual media differed over question types as well. The analysis of the two definition and two procedural questions (n= 887, 271 of which contained visual media) showed that visual media were more frequent with procedural questions than definition questions (χ^2 (1) = 29.23, p< .001). Moreover, the distribution of the functions of visual media differed (χ^2 (2) = 22.70, p< .001). Decorative visuals are overrepresented in answers to definition questions, and underrepresented in answers to procedural questions; informative visuals were underrepresented in answers to definition questions.

Experiment II: Evaluation

Participants and stimuli

Participants were 72 native speakers of Dutch (43 female, 18-64 years old). We selected 16 medical questions for which the production corpus contained: (i) an informative visual, which adds new information to the answer and (ii) a decorative visual, which does not. The set consisted of eight definition questions and eight procedural questions. For each question four answer presentation versions were constructed: a brief and an extended answer, each of which was combined with an informative and a decorative visual (see Figure 1). The brief answer (average 26 words) gave a direct answer to the question, while the extended answer (average 66 words) also provided relevant background information.

Figure 1: *Example of two conditions: a brief answer with a decorative visual (left) and an extended answer with an informative visual (right).*

Design and procedure

The experiment had a 4 (answer presentation) × 2 (question type) mixed factorial design, with answer presentation as between participants variable and question type as within participants variable. After a short practice session, participants studied 16 question-answer combinations, one at a time. After each combination, they were shown the same combination with at the bottom two seven-point Likert scales (implemented as radio buttons) which they had to use to rate the informativeness and the attractiveness of the answer.

Results

Table 3*: Mean results for the informativeness and attractiveness of answer presentation types (ratings range from 1 = "very negative" to 7 = "very positive"; standard deviations in parentheses).*

Factor	Question type	Text with a decorative visual				Text with an informative visual			
		Brief		Extended		Brief		Extended	
Informative?	Definition	3.83	(1.13)	4.01	(1.30)	4.91	(.81)	4.97	(1.20)
	Procedural	3.70	(1.26)	4.27	(1.18)	5.53	(.70)	5.40	(.84)
	Total	3.76	(1.16)	4.14	(1.19)	5.22	(.69)	5.18	(1.00)
Attractive?	Definition	3.93	(.87)	3.76	(1.14)	4.43	(.88)	4.69	(1.01)
	Procedural	4.18	(1.12)	4.18	(1.10)	4.95	(.84)	5.08	(.76)
	Total	4.06	(.96)	3.97	(1.07)	4.69	(.75)	4.89	(.79)

Table 3 shows that brief answers with an informative visual were evaluated as most *informative*, brief answers with a decorative visual as least informative ($F [3,68] = 9.32$, $p < .001$, $\eta^2_p = .29$). Answers to procedural questions were evaluated more informative than to definition questions ($F [1,68] = 15.13$, $p < .001$, $\eta^2_p = .18$). Finally, an interaction was found between answer presentation and question type

(F [3,68] = 4.27, p< .01, η^2_p = .16). This interaction can be explained as follows: for both brief (F [1,17] = 17.12, p< .005, η^2_p = .50) and extended (F [1,17] = 7.31, p< .025, η^2_p = .30) answers with an informative visual, procedural answers were evaluated as more informative than definition answers. For answers with an illustrative visual no significant differences were found between the two question types.
Table 3 also shows that extended answers with an informative visual were evaluated as most *attractive*, while extended answers with a decorative visual were evaluated as least attractive (F [3,68] = 4.64, p< .01, η^2_p = .17). Also, procedural answers were evaluated as more attractive than definition answers (F [1,68] = 20.59, p< .001, η^2_p = .23). No interaction was found between answer presentation format and question type (F<1).

Conclusion

The results of the *production* experiment showed that answer presentations were affected by answer length. Brief answers were accompanied more often by informative visuals, representational visuals were more frequent in extended answers. This is likely caused by the mere fact that it is easier for visuals to be more informative as the text is less extended and thus informative. Answer presentations were also affected by question type. Representational visuals were most frequent in definition answers, and informative visuals in procedural answers.

The results of the *evaluation* experiment showed that question type influenced participants' assessment of the informativeness of text and visual combinations. *Procedural* answers with informative visuals were evaluated as more informative than *definition* answers with informative visuals. An explanation for this could be that medical procedures -as they occurred in this experiment- lend themselves better to be visualized than definitions, because they have a dynamic and spatial character, whereas definitions more often concern abstract concepts that are less easily visualized.

Another interesting result was that brief answers with an informative visual were evaluated as most informative, which is consistent with the result in the production experiment. Interestingly however, extended answers with an informative visual were evaluated as most attractive, which suggests that users like complete information together with high informative visuals.

Future research

The experiments described in this paper offer many opportunities for further work. For example, it would be interesting to investigate whether individual differences, like prior knowledge or learning preferences (i.e., verbal vs. visual) affect participants' assessment on the informativeness and attractiveness of different unimodal and multimodal answer presentations.

The results of the production experiment showed that most answers were produced without using any visuals. A possible explanation for this result could be that the participants could not find a suitable visual on the internet that could be inserted in the PowerPoint presentation. Therefore, it would be interesting to redo the experiment in a more controlled setting in which participants are asked to produce answers based on a predefined corpus of visual media.

Finally, in both experiments offline research methods were used to investigate the role of visuals in multimodal information presentation. The production and evaluation experiment have provided insights on how and when people produce information in a multimodal way. However, what is unclear is how multimodal information presentation is actually processed. Eye tracking could be a useful method to investigate how people process information from different modes and whether different types of multimodal information presentation are processed and integrated differently.

References

Arens, Y., Hovy, E. & Vossers, M. (1993). On the knowledge underlying multimedia presentations. In M. T. Maybury (Ed.), Intelligent Multimedia Interfaces (pp. 280-306). Menlo Park: AAAI Press.

Bernsen, N. (1994). Foundations of multimodal representations. A taxonomy of representational modalities. Interacting with Computers. 6, 347-371.

Carney, R., & Levin, J. (2002). Pictorial illustrations still improve students' learning from text. Educational Psychology Review, 14, 5-26.

Heiser, J., Phan, D., Agrawala, M., Tversky, B., & Hanrahan, P. (2004). Identification and validation of cognitive design principles for automated generation of assembly instructions. Proceedings of Advanced Visual Interfaces, 311-319.

Hooijdonk, C.M.J., van, & Krahmer, E.J. (2008). Information modalities for procedural instructions: the influence of text, static and dynamic visuals on learning and executing RSI exercises. IEEE Transactions on Professional Communication, 51, 50-62.

Mayer, R. (2005). The Cambridge handbook of multimedia learning. Cambridge: Cambridge University Press

Mayer, R., & Moreno, R. (2002). Aids to computer-based multimedia learning. Learning & Instruction, 12, 107-119.

Michas, I., & Berry, D. (2000). Learning a procedural task: effectiveness of multimedia presentations. Applied Cognitive Psychology, 14, 555-575.

Tversky, B., Morrison, J., & Bétrancourt, M. (2002). Animation; can it facilitate? Int. J. Human-Computer Studies, 57, 247-262.