Illustrating Answers: An Evaluation of Automatically Retrieved Illustrations of Answers to Medical Questions

Wauter Bosma and **Mariët Theune**[1]
Charlotte van Hooijdonk and **Emiel Krahmer** and **Fons Maes**[2]

Abstract. In this paper we discuss and evaluate a method for automatic text illustration, applied to answers to medical questions. Our method for selecting illustrations is based on the idea that similarities between the answers and picture-related text (the picture's caption or the section/paragraph that includes the picture) can be used as evidence that the picture would be appropriate to illustrate the answer. In a user study, participants rated answer presentations consisting of a textual component and a picture. The textual component was a manually written reference answer; the picture was automatically retrieved by measuring the similarity between the text and either the picture's caption or its section. The caption-based selection method resulted in more attractive presentations than the section-based method; the caption-based method was also more consistent in selecting informative pictures and showed a greater correlation between user-rated informativeness and the confidence of relevance of the system. When compared to manually selected pictures, we found that automatically selected pictures were rated similarly to decorative pictures, but worse than informative pictures.

1 INTRODUCTION

According to Mayer's [11] well-known *multimedia principle*, people learn better from words and pictures than from words alone. Nevertheless most question-answering (QA) systems, which can automatically answer users' questions that are posed in natural language, still present their answers using a single modality, in the form of text snippets retrieved from a document corpus. Any pictures occurring in the documents are generally ignored, since the text-oriented retrieval methods used in QA systems cannot deal with them. A solution for dealing with non-textual media that has been proposed for use in multimedia summarization and retrieval is to analyze and convert the media content to a semantic representation usable by the system [10, 12, 6, 13]. However, automatic analysis of media content is difficult and often unreliable, while manual annotation is very laborious. Another solution, which according to de Jong et al. [9] is often overlooked, is the use of related linguistic content instead of the media items themselves. If related text adequately describes a media item, text-based retrieval methods can be used to retrieve non-textual media.

Bosma [3] proposed a method for extending the answers returned by a QA-system with appropriate illustrations by searching pictures whose related text is similar to the text of the answer. Pictures are selected by taking the best match of the answer text and a text snippet automatically associated with the picture. This method has been applied in the IMIX system for answering medical questions [5]. The purpose of the IMIX system is to answer medical questions from non-expert users, of the kind to which answers can be typically found in an encyclopedia. Questions can be typed or spoken (in Dutch), and answers are presented using speech, text and pictures. Questions can be asked in isolation, but the system is also capable of engaging in dialogs and answer follow-up questions.

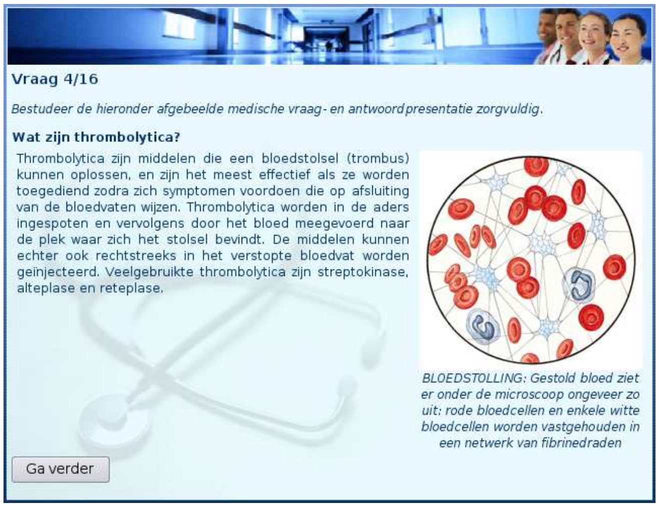
This paper presents a user evaluation of Bosma's [3] picture selection method. In the experiment, answer presentations with automatically selected pictures were rated by naive participants judging the attractiveness and informativeness of the text-picture combination. We also investigated the influence of the different presentations on learning. The experimental design was the same as that used by van Hooijdonk et al. [8], who evaluated manually created answer presentations consisting of different text-picture combinations. We repeated their experiment for answer presentations with automatically retrieved pictures, comparing two versions of the automatic picture retrieval method: one where the picture's textual annotation consists of its caption (resulting in 'caption-selected' illustrations), and one where the annotation is a part of the text near which the picture was found (resulting in 'section-selected' illustrations).

In the following sections, we first explain the picture selection method that is evaluated (Section 2). Then we describe the set-up of the evaluation experiment (Section 3) followed by a discussion of the results (Section 4). We end with some concluding remarks (Section 5).

2 AUTOMATIC TEXT ILLUSTRATION

Our picture selection method is an application of the query-based summarization framework of [4], which is applied in IMIX to generate extended answers consisting of a paragraph-sized text. In QA, the answer's content is drawn from a set of documents (the source documents) which provide an answer but were not necessarily written to answer the query. The query-based summarization approach relies on a combination of one or more feature graphs representing the source documents. The graphs express relations between the documents' content units, and are constructed using information about unit content (e.g. based on cosine similarity) or context (e.g. based on layout) to relate the units. This way, content can be presented

[1] University of Twente, The Netherlands, email: W.E.Bosma@utwente.nl, M.Theune@utwente.nl
[2] Tilburg University, The Netherlands, email: C.M.J.vanHooijdonk@uvt.nl, E.J.Krahmer@uvt.nl, Maes@uvt.nl



Vraag 4/16

Bestudeer de hieronder afgebeelde medische vraag- en antwoordpresentatie zorgvuldig.

Wat zijn thrombolytica?

Thrombolytica zijn middelen die een bloedstolsel (trombus) kunnen oplossen, en zijn het meest effectief als ze worden toegediend zodra zich symptomen voordoen die op afsluiting van de bloedvaten wijzen. Thrombolytica worden in de aders ingespoten en vervolgens door het bloed meegevoerd naar de plek waar zich het stolsel bevindt. De middelen kunnen echter ook rechtstreeks in het verstopte bloedvat worden geïnjecteerd. Veelgebruikte thrombolytica zijn streptokinase, alteplase en reteplase.

BLOEDSTOLLING: Gestold bloed ziet er onder de microscoop ongeveer zo uit: rode bloedcellen en enkele witte bloedcellen worden vastgehouden in een netwerk van fibrinedraden

Figure 1. Example of an answer presentation consisting of text and an automatically selected picture. The presentation answers the question *What are thrombolytics?* The text of the answer explains that thrombolytics are drugs used to dissolve blood clots. The picture depicts a schematic representation of clotted blood.

for which there is just indirect evidence of relevance. For instance, a sentence that is adjacent – and thus contextually related– to a sentence that is similar to the query may be included in the answer, even though it is only indirectly linked to the query.

This concept may also be applied to multimedia. A picture can be related to a piece of text by using layout information. A straightforward relatedness clue of text and picture is when the text is the picture's caption, but also if the picture belongs to a certain paragraph or section, the section and the picture may be considered related. When the relevance of the text is established, the relevance of the picture is established indirectly.

In the IMIX system, this approach is used to select the best picture to illustrate a given textual answer to a medical question. To find this picture, the illustration system compares the text of the answer with picture-associated text. The more similar the two text passages, the more likely the picture is relevant. The picture-associated text is interpreted as a textual representation of the picture. This may be either the picture's caption or the paragraph (or section if no single paragraph could be related to the picture) in which the picture was found. The relevancy of a picture for the answer is calculated as:

$$R_{picture}(i,t) = cosim(t, text(i)) \tag{1}$$

where $R_{picture}(i,t)$ is the relevancy of picture i to text t; and $text(i)$ is the text associated with picture i. The function $cosim(a,b)$ calculates the cosine similarity of a and b.

Cosine similarity is a way of determining lexical similarity of text passages. The idea behind cosine similarity is that a text's meaning is constituted by the meaning of its words. To measure cosine similarity between two passages, we represent both texts as a vector whose elements represent the contribution of a word to the meaning of the passage. Before measuring the cosine similarity, words are stemmed using Porter's stemmer [14]. The cosine similarity is calculated as follows:

$$cosim(a,b) = \frac{\sum_{k=1}^{n} a_k \cdot b_k}{|a| \cdot |b|} \tag{2}$$

where $cosim(a,b)$ is the similarity of passages a and b; n is the number of distinct words in the passages. Both passages are represented as a vector of length n, with a_k representing the contribution of word k to passage a. The denominator ensures that passage vectors are normalized by their lengths. The value $|a|$ is the length of passage vector a, measured as $\sqrt{\sum_{k=1}^{n} a_k^2}$.

Determining how much a particular word contributes to the meaning of a passage is called *term weighting*. In this paper, we use $tf \cdot idf$ term weighting, i.e. the contribution of a word to a passage is calculated as the word's occurrence frequency in the passage (term frequency, TF) multiplied by the word's inverse document frequency (IDF). IDF is a measure of how characteristic the word is for a passage. To measure the inverse document frequency, we require a large set of passages. In this paper, we use the passage vectors of picture-associated text for all pictures in the corpus, plus the passage vector of the answer text. A word occurring in few of these passages receives a high IDF value, because the low occurrence rate makes it descriptive of the few passages it appears in. Conversely, a word occurring in many passages receives a low IDF value. The contribution of word k to passage a is measured as follows:

$$a_k = tf_{a,k} \cdot idf_k \tag{3}$$

where $tf_{a,k}$ is the number of occurrences of word k in passage a; and idf_k is the IDF value of word k. The IDF value is calculated as follows:

$$idf_k = log \frac{|D|}{|\{d \mid d \in D \wedge k \in d\}|} \tag{4}$$

where $|D|$ is the number of passages in the corpus (i.e. the number of pictures plus one); and the denominator is the number of documents which contain the word k.

The final answer presentation consists of the textual answer and the most relevant picture and its caption. An example of an answer presentation containing an automatically selected picture is given in Figure 1 (this is a screen shot showing one of the answer presentations from our experiment, see the next section).

3 THE EVALUATION EXPERIMENT

We carried out an evaluation experiment in which participants evaluated a set of 16 text-picture answer presentations to medical questions. The pictures in the presentations were selected automatically using the method described above. Apart from the pictures used in the answer presentations, the study was identical to the study of manually created presentations by van Hooijdonk et al. [8]. This includes the textual component of the answers. Below we describe the creation of the stimuli used in the experiment, the participants and the experimental procedure.

3.1 Questions and textual answers

In our study, we used the same set of 16 general medical questions that had been used by [8]. Certain properties of the questions in this set were systematically varied, in order to investigate the effect of question type on the effect of the different answer presentations. Of the 16 questions, half were definition questions and half were procedural questions. Of the eight questions in both groups, half referred to body parts and half did not. Table 1 shows examples of the questions used. References to body parts may be indirect, as is the case in the first question in Table 1.

Table 1. Examples of medical questions. Questions are equally divided in the categories of *definition questions* (Def.) or *procedure questions* (Proc.); and in questions which refer to body parts and questions which do not.

Type/Bodypart	Question
Def./Yes	Where is testosterone produced?
Def./No	What does ADHD stand for?
Proc./Yes	How to apply a sling to the left arm?
Proc./No	How to organize a workspace in order to prevent RSI?

For each medical question, van Hooijdonk et al. [8] formulated a concise and an extended textual answer. A concise answer gives a direct answer to the question, and nothing more, while the extended answer also provides relevant background information (c.f. [2]). The average lengths of the concise answers and the extended answers were approximately 26 words and 66 words respectively.

The textual component of an answer presentation was a manually written reference answer. Manual text is used in order to be able to concentrate on evaluating the multimedia aspect – the quality of the text-picture combination. In the experiment reported here, we reused the answer texts from [8] but combined them with new, automatically selected pictures as described below. The text is based on answers produced by a study in which participants answered the same questions as the ones used here, using any information available, including web search. This procedure is described in detail in [8].

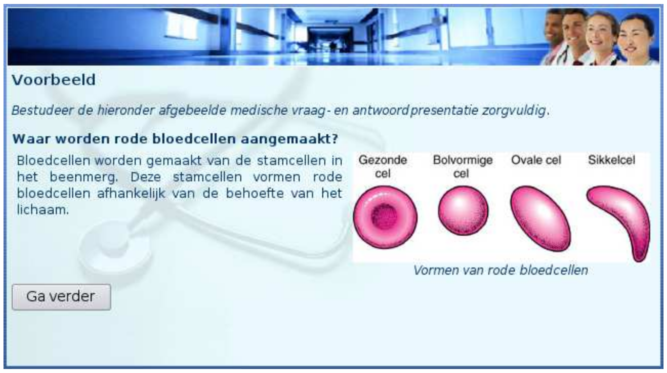
3.2 Illustrating the answers

We created a corpus of annotated pictures to be used for automatically illustrating the textual answers. The pictures as well as their textual annotations were automatically extracted from two medical sources providing information about anatomy, processes, diseases, treatment and diagnosis. Both are intended for a general audience and written in Dutch. The first source, *Merck Manual medisch handboek* [1], Merck in short, contains 188 schematic illustrations of anatomy and treatment, process schema's, plots and various types of diagrams. The other source, *Winkler Prins medische encyclopedie* [7], WP in short, contains a variety of 421 pictures, including photographic pictures, schema's and diagrams. These sources were selected because they cover the popular medical domain and they are relatively structured – paragraph boundaries are marked in the text and all 609 pictures have captions.

In this experiment, for each of the textual answers, two presentations were generated by illustrating them using the algorithm described in section 2, applied to the picture corpus described above. For one of the presentations for each answer, the picture's caption was used as associated text. For the other presentation the picture was associated with the smallest unit of surrounding text from its original document; this could be a section or a paragraph. The surrounding text was extracted automatically, using meta-information in the document such as XML tags.

The average distribution of selected pictures from our two sources (Merck 33 percent; WP 66 percent) reflects the distribution in our picture corpus (Merck 31 percent; WP 69 percent). Table 2 lists the number of selected pictures from each source for the four selected conditions, with percentages given between brackets. Note that for each condition, 17 pictures were selected: 16 for the answer presentations to be evaluated, plus one for an example presentation that was presented to the participants (see Section 3.4).

The corpus did not contain an appropriate picture for all answers, which forced the illustration system to select less appropriate pictures for some of the presentations. In some cases the selected picture was plain irrelevant, but in some other cases, the picture was related to the text but had a different perspective. For instance, the picture in Figure 2 addresses the deformation of red blood cells rather than their generation. This problem may have been augmented by the fact that the pictures in our corpus have a high information density; only few pictures have a decorative function only (i.e., they do not add any information to the related text). Consequently, the pictures are relatively specific to their original context, which complicates their reuse in a slightly different context.



Voorbeeld

Bestudeer de hieronder afgebeelde medische vraag- en antwoordpresentatie zorgvuldig.

Waar worden rode bloedcellen aangemaakt?

Bloedcellen worden gemaakt van de stamcellen in het beenmerg. Deze stamcellen vormen rode bloedcellen afhankelijk van de behoefte van het lichaam.

Figure 2. Example of a picture which is related but not complementary to the answer text. The presentation answers the question *Where are red blood cells generated?* The text explains that red blood cells are generated from stem cells in the bone marrow. Rather than illustrating this, however, the picture shows various deformations of red blood cells.

Table 2. Number of pictures (with percentages in brackets) selected from Merck [1] and WP [7].

Condition	Merck	WP
Brief text; caption-selected picture	6 (35%)	11 (65%)
Extended text; caption-selected picture	4 (24%)	13 (76%)
Brief text; section-selected picture	6 (35%)	11 (65%)
Extended text; section-selected picture	7 (41%)	10 (59%)

The answer presentations were created as a web page headed by the question (in bold face), followed by the answer text on the left and the best-matching picture on the right side of the page. Regardless which method had been used to select the picture (caption-based or section-based), we considered the caption part of the picture and thus presented it along with the picture in the answer presentation. Since all pictures in our corpus had a caption, this was always included. If the text surrounding the picture had been used for its selection, this text was not included in the answer presentation.

A complicating factor here was that captions vary greatly in length, especially in the WP corpus. Table 3 shows details of the distribution of caption lengths (for comparison, details about section lengths are given in Table 4). The most extreme case was a caption as long as 428 words. Since the textual component of the answer presentations averaged only 26 or 66 words (for concise and extended presentations respectively), presenting very long captions along with the pictures would lead to an imbalance between the amount of text in the caption and the amount of text in the textual component of the answer. In order to prevent excessive caption lengths, in the answer presentations the captions were truncated to their first sentence. So only the caption's first sentence was presented along with the picture, rather than the caption as a whole. This was done *after* picture selection, so it did not affect the picture selection process.

Table 3. Caption length statistics of the Merck corpus [1] and the WP corpus [7].

	Caption length (words)	
	Average	SD
Merck	4.4	1.9
WP	39.1	42.9
Combined	28.4	39.1

Table 4. Section length statistics of the Merck corpus [1] and the WP corpus [7].

	Section length (words)		
	Average	SD	range
Merck	354	325	[30,1944]
WP	67	48	[5,336]
Combined	156	227	[5,1944]

3.3 Participants

Seventy five people participated in the experiment: 44 female and 31 male, between 18 and 55 years old. Fifty six of them (75 percent) were students recruited from Tilburg University. The remaining 25 percent were recruited from various e-mail lists. None had participated in the experiments of [8]. The participants were randomly assigned to one of the four conditions (concise or extended text, selection by means of caption or surrounding text), of which they were shown all 16 answer presentations.

3.4 Experimental procedure

The participants were invited by e-mail to participate. This e-mail shortly stated the goal of the experiment, the amount of time it would take to participate, the possibility to win a gift certificate, and the URL of the experiment. The experiment, created using WWStim [15], was entirely online.

When the participants accessed the experiment, they first received instructions about the procedure. The participants were told that they would receive the answer presentations of 16 medical questions, which they would have to study carefully and then assess their informativeness and their attractiveness. Next, the participants entered their personal data, i.e., age, gender, level of education, and optionally their e-mail to win a gift certificate.

After participants had filled out their personal data, they practiced the procedure of the actual experiment in a practice session: they were given the medical question *Where are red blood cells produced?*. First, the participants answered on a seven-point Likert scale how confident they were to know the answer to this medical question. Subsequently, the participants were shown the answer to the medical question corresponding to the condition they were assigned to. (See Figure 2 for the concise-answer, caption-selected picture condition.) The participants studied the answer presentation until they thought that they could assess its informativeness and attractiveness. Then, the participants were shown the medical question, the answer presentation, and a questionnaire. This questionnaire consisted of five questions, asking them to rate on a seven-point Likert scale:

1. the clarity of the text;
2. the informativeness of the answer presentation;
3. the attractiveness of the answer presentation;
4. the informativeness of the combination of text and picture;
5. the attractiveness of the combination of text and picture.

The participants judged the informativeness of the text-picture combination instead of directly assessing the relevance of the picture. This is because the experiment in [8] contained manually selected pictures only, for which relevance was assumed (although a distinction was made between decorative and informative pictures). In contrast, automatic pictures may be irrelevant or somewhat relevant. However, we chose not to change the design of the experiment in order to get comparable results. (See Section 4.3 for a comparison between presentations with manually and automatically selected pictures.)

After completing the practice session, the participants started with the actual experiment, proceeding in the same way as during the practice session. When they were finished with their assessment of the answer presentations to the 16 medical questions, the participants received a post test which was the same for all participants (regardless the experimental condition). In the post test, the participants had to answer the same 16 questions of which they had rated the answer presentations in the previous part of the experiment. This was done in the form of a multiple choice test, in which each medical question was provided with four textual answer possibilities. Of these four answer possibilities, one answer was correct and the other three were plausible incorrect ones. The order in which the medical questions were presented in the post test was the same as in the actual experiment. Note that – with respect to the concise textual answer – the additional information in the extended textual answers and in the pictures was not necessary to answer the question in the post test correctly.

4 RESULTS

The results of the assessments were normalized to be in the range $[0..1]$. A rating n between one and seven (inclusive) was normalized as $\frac{1}{6}(n-1)$.

For processing the results, we used the following, non-standard method. For each condition and each medical question and assessment question, we calculated the average assessment. For pair-wise significance testing of differences between two experimental conditions for a particular assessment question, we measured the percentage of answer presentations for which the rating of one condition was higher than that of another. A condition that consistently received higher average ratings than the other for each medical question got a score of 100 percent; consequently, the other condition got a relative score of 0 percent. Significance is tested by means of 10^6-fold approximate randomization. A difference is considered significant if the null hypothesis (that the sets are not different) could be rejected at a certainty greater than 95 percent ($p < 0.05$), unless stated otherwise.

The reasons for using the mutual rank instead of the average judgment are as follows. To see if one type of answer presentation is better than another, one could simply check whether the difference in average scores is significant. However, while a single average score is useful as a rough quality indication, it may not be the best method for a pairwise comparison.

If the difference in scores between two types of answer presentation does not tell anything about the difference in quality other than which one is better, a comparison can have only three possible outcomes: one is better, the other is better, or their quality is equal. If this is accepted, it remains to be seen whether the score averages are reliable for significance testing. The standard deviation of ratings of answers to some medical questions was higher than the standard deviation for answers to other medical questions. As a result, some

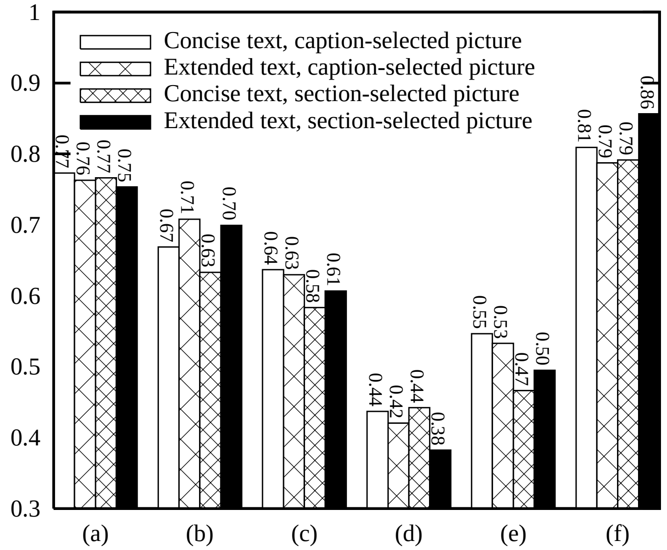


Figure 3. Average assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test.

medical questions affect the average rating more than others. This makes it less likely to find significant differences in rating. Using the mutual rank avoids this problem.

4.1 Caption or section?

Figure 3 shows an overview of the average assessments per condition. The level of clarity of the textual component of the answer (Figure 3 (a)) was judged similar. No significant differences between any two conditions were found.

Regarding the informativeness of the answer presentation as a whole (Figure 3 (b)), extended answers were rated significantly more informative than concise answers. However, for extended answers, the combination of picture and text (Figure 3 (d)) was judged less informative. This effect was the strongest for pictures that were selected using their surrounding section, although the differences were not significant.

The presentation (Figure 3 (c)) as well as the picture/text combination (Figure 3 (e)) was rated significantly more attractive if the pictures were selected based on their captions than if they were selected based on their surrounding section. The attractiveness of the presentation or the picture/text combination was not affected by the length of the textual component of the answer.

All in all, the presentations containing a section-selected picture were less informative and less attractive than the presentations containing a caption-selected picture. Apparently, captions are more representative of the content of a picture, and thus are more reliable indicators of the picture's relevance to the answer text. This is not entirely surprising, as the content of a caption generally describes (only) the picture, whereas the text surrounding a picture may also contain unrelated content.

In seeming contradiction with the good ratings of caption-selected pictures, in the post test where participants had to select the correct answer in a multiple choice test, participants who were shown section-selected pictures gave significantly more correct answers than other participants when the section-selected picture was included in a presentation with an extended textual component. This is a remarkable result because these pictures were rated least informative. A possible explanation for this is that the participants concentrated less on the picture (because they quickly dismissed it as less relevant) and more on the text. After all, the information in the picture was not required to answer the questions in the post test.

4.2 The value of confidence

The selection criterion for automatic pictures was the cosine similarity of the textual component of the answer and the text associated with the picture (a caption or a section, depending on the condition). The picture with the highest cosine similarity was selected. Because cosine similarity is used as a measure of relevance, this value can be interpreted as a *confidence value*, i.e. how confident the system is that the selected picture is actually relevant. If the cosine similarity is actually a good indicator of relevance, one would expect a high correlation between cosine similarity and relevance. In the IMIX system, in which this picture selection method is implemented, the answer is presented text-only if no picture has a confidence (cosine similarity) above a certain (configurable) threshold. Table 5 shows the averages of the cosine similarity values of the pictures selected for the answers in the experiment described in this paper.

Table 5. Statistics of the cosine similarity of the textual component of the answer and the text passage used for indexing the selected picture.

Condition	Average	(standard deviation)
Brief text; caption-selected picture	0.190	(0.00788)
Extended text; caption-selected picture	0.188	(0.00631)
Brief text; section-selected picture	0.133	(0.00501)
Extended text; section-selected picture	0.162	(0.00654)

But what is the meaning of cosine similarity as a confidence value? Cosine similarity can be used to predict the relevance of the picture if there is a correlation between the cosine similarity and the experimental participants' judgments of a presentation. Figure 4 shows the correlation of the confidence (cosine similarity) value and the participant judgments. A value of 1 (or -1) indicates a perfect increasing (or decreasing) linear correlation. This correlation was greatest for the participant judgments of the informativeness of the text-picture combination (0.51 and 0.44 with concise and extended text respectively). This is an encouraging result, given that this aspect seems to correspond most closely to picture relevance. With respect to attractiveness, the correlation with confidence was significantly greater for concise answers than for extended answers. There was only a slight difference in correlation between attractiveness and confidence for different picture selection methods.

Remarkably, participants perceived the textual component of the answer as less clear when the confidence value of the picture was greater. This puzzling result suggests that relevant pictures negatively affect the clarity of the answer text rather than enhance it. A possible explanation is that any mismatches between picture and text may be more confusing when text and picture seem closely related than when the picture obviously does not fit the text, in which case it can be easily ignored and does not influence the interpretation of the text.

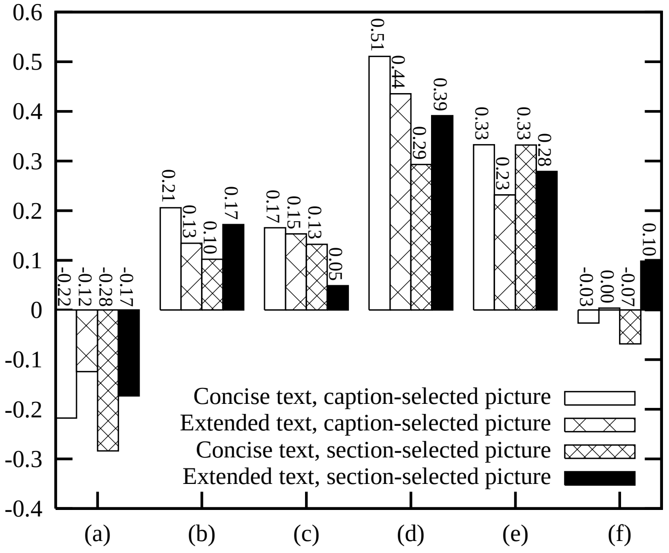


Figure 4. Pearson correlation coefficient between the confidence of picture selection and the assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test.

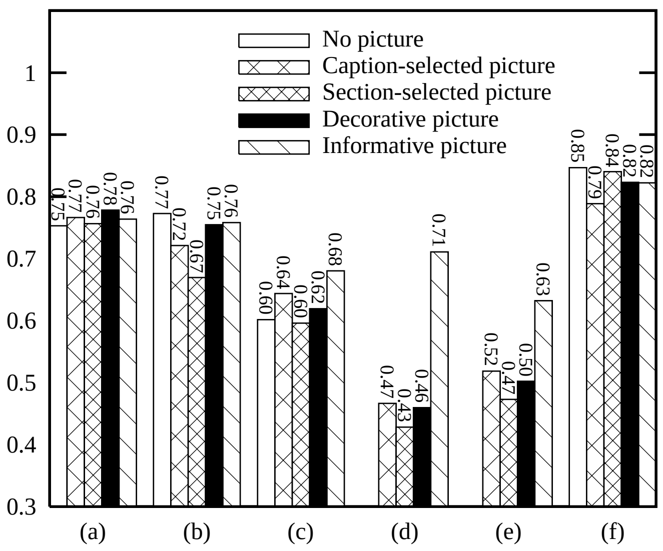


Figure 5. Average assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test. For comparability, these results include only registered students from Tilburg University. Therefore, the actual values may differ slightly from Figure 3.

4.3 Automatic or manual?

As mentioned earlier, apart from the answer presentations themselves, the design of the experiment was identical to the experiment described in [8]. This allows us to compare the evaluation results of our automatically illustrated answer presentations to those of [8], who evaluated manually created answer presentations.

In the experiment of [8], the answer presentations consisted of the same (concise or extended) textual component used in the current experiment, in combination with either no picture, a decorative picture, or an informative picture (i.e. six experimental conditions in total). These manually selected pictures can be regarded as a *gold standard* for decorative and informative pictures respectively. However, in practice, it is unlikely that this gold standard can be achieved with the set of 609 medical pictures used for automatic picture selection in our experiment, because the picture sources used by [8] were unrestricted and thus offered far more opportunities to find a suitable illustration for a given answer text.

A large portion of participants in both experiments were students from Tilburg University. Because these students received course credits for participation, they filled in their student registration number, which made it possible to distinguish them from other participants. However, in both experiments, other participants took part from outside this community, and we found significant differences between the registered students and the other participants with respect to their answers to some of the assessment questions. On average, for 65 percent ($p < 0.001$) of the answer presentations of [8], the informativeness of the presentation was rated higher by student participants than by other participants. In the same experiment, students rated the text-picture combinations more informative (60 percent, $p < 0.001$) and less attractive (58 percent, $p < 0.01$) than other participants. The answers to other assessment questions were similar for both groups, or slightly different.

The results of two experiments are comparable only if the group of participants in one experiment is similar to the participants of the other experiment. However, students and non-students are shown to produce different results, rendering the participant groups as a whole dissimilar. Therefore we filtered the non-students out from our comparison between automatically and manually selected illustrations, to ensure that the experimental conditions are the only variables over both experiments. Since the students participating in both experiments were recruited within a short time frame using the same communication channels, we consider both groups as fully comparable.

In total, 98 participants (70 female, 28 male) in both experiments were registered students. Of them, 42 contributed to the experimental conditions of [8] and 56 contributed to the conditions from our experiment, described in section 3. No one participated twice. The average assessments of the 98 participants are shown in Figure 5. These results combine the 16 concise and the 16 extended answer presentations, comprising 32 data points for each condition and assessment question.

The informativeness of text-picture combinations as well as the attractiveness of the presentation was similar when the answer contained an automatically selected picture, a manually selected decorative picture, or no picture at all. No significant differences were found. However, the text-picture combination of manually selected informative pictures was rated significantly more informative than the text-picture combination of manually selected decorative pictures and automatically selected pictures. Answer presentations were rated significantly less informative if the presentation contained a section-selected picture than if the answer contained an informative picture, a decorative picture, or no picture at all. Presentations containing caption-selected pictures are not significantly less informative than presentations with informative pictures.

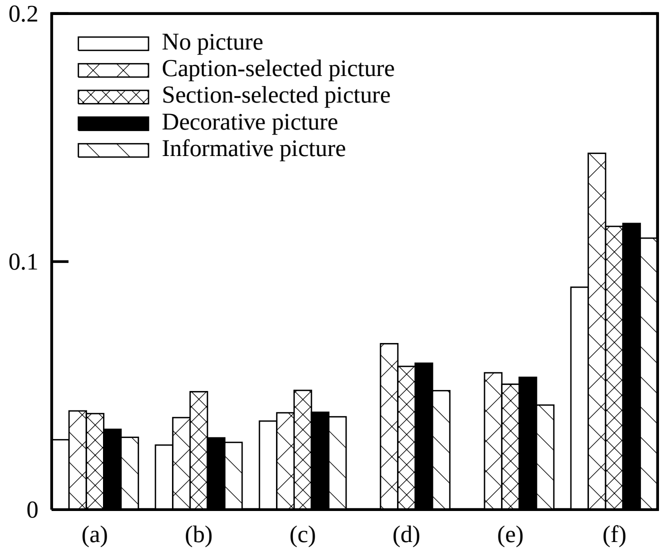


Figure 6. Standard deviations per answer presentation in the assessments of (a) textual clarity; (b) informativeness of the presentation; (c) attractiveness of the presentation; (d) informativeness of the text-picture combination; (e) attractiveness of the text-picture combination; and (f) the average percentage of correct answers in the post test. For comparability, these results include only registered students from Tilburg University.

Average ratings of automatic presentations may have been negatively affected by inconsistent performance of the picture selection algorithm. In some cases, the algorithm selected an irrelevant or a somewhat irrelevant picture because there was no appropriate picture in the database or simply because the algorithm failed to find it. If the relevance of automatic pictures is less consistent than that of manual pictures, this should reflect in the variability of the results. Figure 6 shows the standard deviations of assessments. For automatic pictures, participants indeed show greater variability than for manual pictures in their assessments of textual clarity, informativeness and attractiveness of the answer presentation. Remarkably, we found that the standard deviation of the number of correct answers in the post test was also greater for pictures which are selected by their captions.

5 CONCLUSION

This paper presented an algorithm for automatic illustration of answers to medical questions in Dutch. It is used in the IMIX question answering system to add appropriate illustrations to textual answers. To evaluate the algorithm, we conducted an experiment, following the same procedure as [8] to evaluate different types of answer presentations on informativeness, attractiveness and influence on learning.

In our experiment, the answer presentations contained a textual and a visual component, of which the text was given and the visual was automatically retrieved from an offline picture database containing 609 pictures. The pictures were automatically extracted from *Merck Manual medisch handboek* [1] and from *Winkler Prins medische encyclopedie* [7]. To find an appropriate picture, the pictures were indexed by a passage of text from the document in which they were found. Two different indexing methods were compared in the experiment, either using the picture's caption for picture selection, or using the section or paragraph that contained the picture. Both selection methods were tested in combination with a concise or an extended textual answer.

Due to limitations of the corpus (i.e. for several questions it did not contain a relevant picture at all) the standard deviations of our results are quite high, which makes it difficult to make any general claims based on them. However, some tentative conclusions can be drawn.

The results indicate that the caption-based picture selection method results in more informative and attractive presentations than the section-based method, although the difference in informativeness was not significant. Furthermore, caption-based picture selection shows a greater correlation between confidence and informativeness, which indicates that the confidence value better predicts the informativeness of the picture. A system could use this to respond by not offering any picture if no relevant picture is available (as is currently done in the IMIX system). All in all, the caption-based picture selection method offers more promising results than the section-based selection method.

An investigation of the relation between system confidence and our experimental results revealed an intriguing negative correlation between textual clarity and the predicted relevance of the selected illustration. Apparently, seeing an answer text in combination with a picture that is related to it, but not fully attuned to it, may be confusing to the user. Problems like these might be solved by the development of post-processing methods to adapt the textual and visual components of the answer presentation to each other, so that they form a more coherent whole.

When compared to manually created answer presentations, we found that answer presentations with an automatically selected picture were largely rated at the same level as presentations with a manually selected decorative picture (which did not add any information to the answer) or even no picture at all. This may be partially explained by the design of the experiment, where the visual element of the answer presentations was not needed to answer the question (since the textual element contained all the required information). Also, the results were undoubtedly influenced by the fact that our picture corpus did not contain appropriate pictures for all answers, in which case the algorithm had no choice but to select an irrelevant picture. To measure the extent of this influence, we should perform a sub-analysis on those questions for which the corpus did contain at least one appropriate picture. In general, we can say that, given the limitations of our corpus, achieving comparable ratings to manually selected decorative pictures is not a bad result.

ACKNOWLEDGEMENTS

This research was performed in the IMOGEN (Interactive Multimodal Output GENeration) project. IMOGEN is a project within the Netherlands Organisation for Scientific Research (NWO) research programme on Interactive Multimodal Information eXtraction (IMIX). The authors would like to thank Jurry de Vos and Lennard van de Laar for their help in setting up the experiment.

REFERENCES

[1] *Merck manual medisch handboek*, eds., Robert Berkow, Mark H. Beers, and Andrew J. Fletcher, Bohn Stafleu van Loghum, Houten, the Netherlands, 2nd edn., 2005.

[2] Wauter Bosma, 'Extending answers using discourse structure', in *Crossing Barriers in Text Summarization Research*, eds., Horacio Saggion and Jean-Luc Minel, pp. 2–9, Shoumen, Bulgaria, (September 2005). Incoma Ltd.

[3] Wauter Bosma, ‘Image retrieval supports multimedia authoring’, in *Linguistic Engineering meets Cognitive Engineering in Multimodal Systems*, eds., E.V. Zudilova-Seinstra and T. Adriaansen, ICMI Workshop, pp. 89–94, Trento, Italy, (October 2005). ITC-irst.

[4] Wauter Bosma, ‘Query-based summarization for question answering’, in *Computational Linguistics in the Netherlands 2004: Selected papers from the fifteenth CLIN meeting*, eds., Ton van der Wouden, Michaela Poß, Hilke Reckman, and Crit Cremers, number 4 in LOT Occasional Series, pp. 29–44, Utrecht, the Netherlands, (2005). LOT.

[5] Lou Boves and Els den Os, ‘Interactivity and multimodality in the IMIX demonstrator’, in *IEEE International Conference on Multimedia and Expo*, pp. 1578–1581, Amsterdam, (July 2005). IEEE Computer Society.

[6] Kees van Deemter and Richard Power, ‘High-level authoring of illustrated documents’, *Natural Language Engineering*, **2**(9), 101–126, (2003).

[7] *Winkler Prins medische encyclopedie*, eds., Peter Fiedeldij Dop and Simon Vermeent, Spectrum, 3rd edn., 1974.

[8] Charlotte van Hooijdonk, Jurry de Vos, Emiel Krahmer, Alfons Maes, Mariët Theune, and Wauter Bosma, ‘On the role of visuals in multimodal answers to medical questions’, in *Proceedings of the 2007 Conference of the IEEE Professional Communication Society*. IEEE, (2007).

[9] Franciska de Jong, Thijs Westerveld, and Arjen de Vries, ‘Multimedia search without visual analysis: the value of linguistic and contextual information’, *IEEE Transactions on Circuits and Systems for Video Technology*, **17**(3), 365–371, (2007).

[10] Mark Maybury and A. E. Merlino, ‘Multimedia summaries of broadcast news’, in *1997 IASTED International Conference on Intelligent Information Systems*. IEEE, (1997).

[11] Richard Mayer, *The Cambridge handbook of multimedia learning*, Cambridge University Press, Cambridge, 2005.

[12] Katashi Nagao, Shigeki Ohira, and Mitsuhiro Yoneoka, ‘Annotation-based multimedia summarization and translation’, in *Proceedings of the 19th international conference on Computational linguistics*, pp. 1–7, Morristown, NJ, USA, (2002). Association for Computational Linguistics.

[13] Valery Petrushin, *Introduction into Multimedia Data Mining and Knowledge Discovery*, 3–13, Springer London, 2007.

[14] M.F. Porter, ‘An algorithm for suffix stripping’, *Readings in information retrieval*, 313–316, (1997).

[15] Theo Veenker. WWStim: A CGI script for presenting webbased questionnaires and experiments, 2005. Website: http://www.let.uu.nl/Theo.Veenker/personal/projects/wwstim/doc/en/.