

BORIS VAN SCHOOTEN
RIEKS OP DEN AKKER
University of Twente
Human Media Interaction
Enschede, The Netherlands
{b.w.vanschooten | h.j.a.opdenakker}@ewi.utwente.nl

Multimodal follow-up questions to multimodal answers in a QA system

Abstract

We are developing a dialogue manager (DM) for a multimodal interactive Question Answering (QA) system. Our QA system presents answers using text and pictures, and the user may pose follow-up questions using text or speech, while indicating screen elements with the mouse. We developed a corpus of multimodal follow-up questions for this system. This paper describes a detailed analysis of this corpus, and its impact on the implementation of our system.

We found that users pose two major types of follow-up question: regular questions, which may be reformulated in such a way as to be answerable by a QA system, and questions asking about a specific picture. We found that users often indicate screen elements with the mouse, even in cases where it may be considered redundant, and that these mouse gestures appear to have a close correspondence to regular anaphors in the utterance. We also found that users use a limited number of ways to indicate screen elements with the mouse. We argue that our QA system will need to annotate its pictures with information about the visual elements that the picture is made up of. This enables appropriate anaphor resolution and answering identity questions about these elements. We present our first results of follow-up question handling and deictic reference resolution, using annotations we made for the pictures of the corpus.

1 Introduction

This paper is part of the IMIX project, which concerns the development of a multimodal question answering (QA) dialogue system answering layman's questions in the medical domain. This paper focuses on dialogue management issues inherent in multimodal interaction, that is, the system presents multimodal answers in the form of text and pictures, and the user asks multimodal follow-up questions in the form of text/speech plus pointing at screen elements using the mouse or a touch screen. While the research on the issues of introducing dialogue capabilities to QA systems is slowly getting off the ground (see for example Van Schooten & Op den Akker (2006) and Galibert et al. (2005)), introducing multimodality in QA is still almost unheard of. In this paper we try to explore this new area in some directions. Multimodality in QA is treated in a broader context in a companion paper (Theune et al., 2006). We shall assume that we have a QA system that is able to produce answers with pictures, either taking the pictures from the document where the text came from, or combining text and pictures from different databases. The domain in which we operate is the medical domain. Nevertheless, there seems to be ample opportunity for generalisation to other domains.

In order to get some data on how users would react multimodally to a multimodal answer, we collected a corpus of multimodal follow-up questions, which we shall analyse in detail in this paper. This corpus was collected by means of a set of canned questions with multimodal answers, rather than a real QA dialogue system. Instead of posing a free-form initial question, the user just selects a question from this set. The user is then presented with the answer, and that's when user can pose an actual, free-form multimodal follow-up question. We asked the users to pose *multimodal* follow-up questions, rather than any kind of follow-up question, in order to arrive at an appropriately large number of multimodal utterances for analysis. We defined "multimodal" primarily as *referring to pictures in the answer*, with or without using pointing gestures. This has the disadvantage of artificiality, which means that the results may be biased in unknown ways, which is not a fatal flaw, since we are primarily interested in the range of phenomena that may occur, rather than their precise relative frequencies.

The users' output modalities were typed text and pointing with the mouse. The users were computer science students and employees, which could access the experiment through a Web page. Before the start of the experiment, the users were presented with one page of instructions, several examples, plus a short introduction to the user interface. The number of available canned question-answer pairs was limited to 20, because these had to be written manually. To make the multimodal aspect of the answers significant, particularly interesting and complex medical pictures and diagrams were chosen. We collected 202 of such multimodal "second questions" from 20 users. Figure 1 shows an impression of the collected data we worked with in this research. Note that the text is in Dutch, so it's not readable.

We will argue that handling multimodal follow-up questions involves several different aspects: handling follow-up questions, recognising the user's gestures and relating them to screen elements, and representing the meaning of pictures. We will discuss each of these aspects in detail in the following sections.

Figure 1. Example interactions from the corpus. At the top is the original question, in the middle is the answer, at the bottom is the user's follow-up question. The encirclings in black are the mouse strokes made by the user. The stippled boxes are the words' bounding boxes. **Left:** original question: "What causes vertigo?", follow-up question: "What do these letters mean?" **Right:** original question: "How do anti-depressants work?", follow-up question: "Why is the serotonin in a kind of sac?"

2 Handling follow-up questions

We classified the utterances in the corpus by looking at the way they may be handled by a QA system. First, we have to tell something about how a QA system may handle follow-up questions. The general issue of handling follow-up questions is very broad, and there does not really exist a set of proven practices. Basically though, a follow-up question is usually handled in a similar way as a stand-alone QA question. The QA system finds a set of documents and images that may contain answers to the question, then an answer is assembled from combining text fragments and images (in IMIX, we only select one text fragment from one document, and possibly add one appropriate image to it). The only difference between a follow-up question and a stand-alone question is that extra context information is taken into account in this answering process. This may be done by either rewriting the question (i.e. Fukomoto et al., 2004) or by passing extra search criteria to the underlying search system (such as a set of topic keywords, i.e. De Boni & Manandhar, 2005). Both amount to the same thing (adding appropriate linguistic context), but rewriting is the most complete, in that a successfully rewritten human-readable sentence implies that we have correctly applied context. Rewriting is, however, not easy. Our experience is that the most fruitful means of rewriting is anaphor substitution. For example, the user asks 'What is the function of this capsule?' while pointing at a certain visual element (a Bowman's capsule), which may be rewritten to: 'What is the function of the Bowman's capsule?'. Anaphor substitution requires recognition of the opportunity, of the anaphor to be substituted, and determination of the referent.

However, not all follow-up questions can be handled in the aforementioned manner; there is also a class of questions which may be considered "non-QA" questions. Among these are questions which concern the specific discourse found in the answer that is reacted to. These can only be answered in the context of that specific answer. Such discourse-related questions are found in unimodal QA, for example, people asking for a missing discourse entity that was accidentally excluded from the text fragment. More complex discourse-related questions also occur, such as people asking about very specific assumptions, statements, or ambiguities in the answer. Understanding such questions may be considered outside of the capabilities or even the philosophy of current QA systems, and an appropriate reaction may be to show more of the document that the answer came from when the system detects a discourse-related question.

In multimodal QA, we found that "non-QA" questions occur more often than in unimodal QA. The most common type is asking for the identity of a visual element. That is, users say something like 'What is this?', or 'Is this the?' while indicating a visual element. Other kinds of visual "discourse" related questions also occur, for example, 'Of what side of the head is this picture?'; 'Where in the picture is the medicine?'; 'In what direction do these flows go?'. Following this line of thought, we classified the follow-up questions in our corpus into different types:

Self-contained: the question does not need any rewriting to make it answerable by a QA.

Regular-rewritable: a regular QA question, which is rewritable using anaphor substitution to form a self-contained question. A dialogue manager may handle this kind of question by detecting which transformation is applicable and finding the appropriate referents and inserting them.

Regular: a question that can be (manually) reformulated so as to form a self-contained QA question. While not rewritable like regular-rewritable, these questions can be handled by a QA in the regular manner.

Visual-element-identity: not answerable without relating to the answer's specific discourse, but answerable by just naming the identity of a particular visual element of a picture in the last answer.

Visual-property: not answerable without relating to the answer's specific discourse, and has something to do with the content of a particular picture, other than visual-element-identity. This is a difficult type of question to handle properly, but might simply be answered by producing a suitable figure caption paraphrasing the entire figure.

We found that 19% of the questions are not multimodal (mostly regular questions that did not include mouse pointing and did not refer to any visual referent), or are not follow-up questions (these are mostly remarks). We discard these in the results presented here. Everything else was classifiable into the five above classes. The relative frequency of these different classes is illustrated in figure 2.

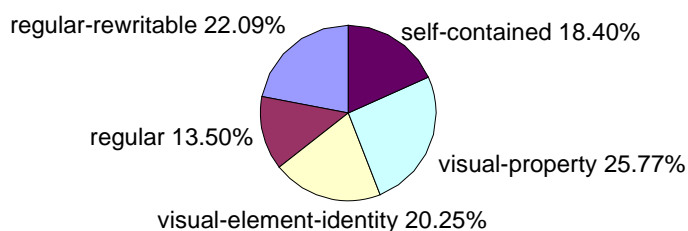


Figure 2. Pie chart showing the percentage distribution of the follow-up question classes in the corpus.

3 Resolving visual referents

In our corpus, there were no follow-up questions which did not have a linguistic component, that is, users never asked a question by means of just pointing. In fact, almost all follow-up questions can be considered primarily linguistic, and the meaning of pointing within the interaction can generally be understood as just hints (though sometimes essential ones) to disambiguate the anaphors and other references in the question text. Referents in the utterances were usually visual elements, but some of the utterances referring to visual elements did not include pointing actions. Instead, the visual elements were typically referred to by means of their colour, shape, or name. Often, a redundant combination of these were used; for example, one user asked 'What function do these blue spots have?' while encircling several blue circles, thus combining colour, shape, and pointing action as hints to disambiguate the utterance's referents. Overall, our findings indicate that traditional (anaphor) reference resolution is a meaningful and important first step in the interpretation of our multimodal utterances. In the rest of this section, we shall try to find out in what ways users refer to visual referents.

How do users indicate using the mouse? We encouraged the use of encircling by producing several encircling examples before the users started the experiment. We consider encircling the most important type of indication, because it allows indication of both location and size of a visual element. However, as is usual in "natural" dialogue systems, we allowed users to use any other kind of pointing action, and users did commonly use several other types of pointing action. Figure 3 shows some examples of these pointing gestures.

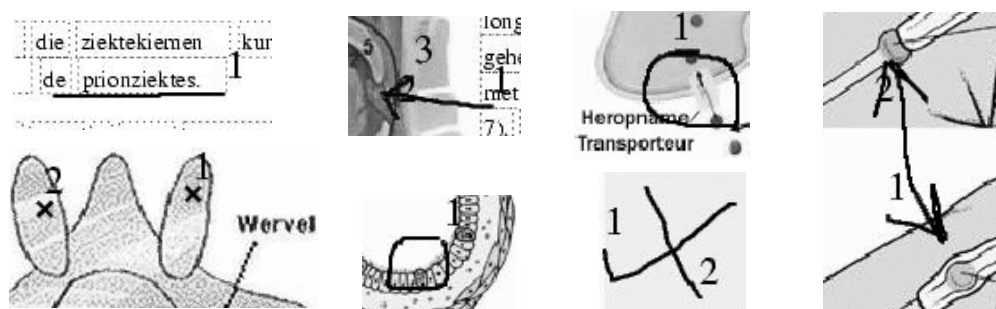


Figure 3. Examples of pointing gestures from the corpus. The strokes are shown in black, with the numbers indicating their order in the utterance. The small crosses indicate tap strokes.

In order to provide a more systematic analysis, we looked at three aspects of the user utterances: the pointing gestures, the anaphors, and the possible relations between these two.

The first aspect involves segmenting the series of mouse strokes of one utterance into a set of pointing gestures. We define a *mouse stroke* as a continuous line or curve drawn with the mouse, and a *pointing gesture* as a set of mouse strokes that has the goal of indicating a visual element. We found that most pointing gestures consist of only one mouse stroke, but some, like arrows, typically consist of two or three mouse strokes. We found that almost all mouse strokes were clearly identifiable as being part of pointing gestures. We found that 81% of the utterances contained at least one pointing gesture, and 23% of these (19% of all utterances) contained multiple ones.

We assigned a type to each pointing gesture. We found that 4 distinct types were sufficient to cover almost all cases: **encircle**, **tap** (that is, just a mouse click), **underline** (as in, underlining a word), and **arrow**. What was left over was a small amount of mouse strokes that appear to be erroneous (labeled as **error**), and a small miscellany category called labeled **other**. Figure 4 shows the relative frequencies of the different classes.

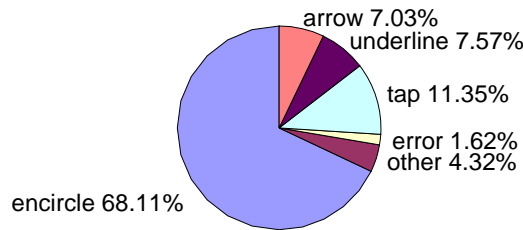


Figure 4. Pie chart showing the percentage distribution of pointing gesture types.

The second aspect we look at is the relation between the pointing gestures and the linguistic components of the questions. This was done by labeling the anaphors that clearly referred to visual elements, and the anaphors that clearly correspond with pointing gestures. We found that there was only a small minority of pointing gestures (9%) for which no anaphor could be pointed out, indicating that anaphors and pointing gestures are closely related. We found no gestures that correspond to multiple anaphors, but some anaphors do refer to multiple gestures. We found that these anaphors were significantly more often in plural form, explicitly indicating a set of referents. Both singular and plural forms were found in all cases, however (see figure 5).

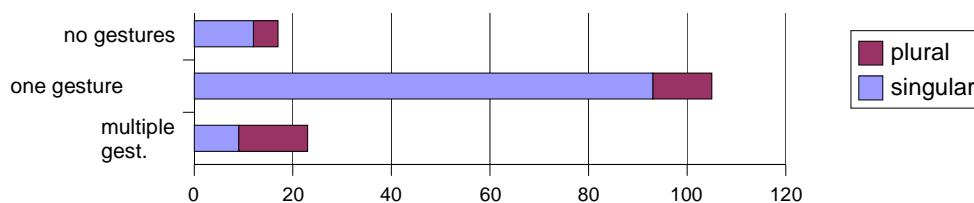


Figure 5. Bar chart showing the anaphor-gesture relationship: the number of gestures that each anaphor corresponded with (either none, one, or multiple), and the plurality of the anaphor.

The third aspect we annotated is the ways in which the identified anaphors provide hints towards resolving their referents, beside the pointing gestures. We classify hints according to the aforementioned types: **colour**, **shape**, and **name**. We found that 49% of the anaphors provide no hints, they were just indications like "this" or "this area". Of the remaining 51%, name occurred the most often by far. Colour and shape were relatively often used simultaneously. As one might expect, name was almost never used simultaneously with colour and shape. Our findings are summarised in figure 6.

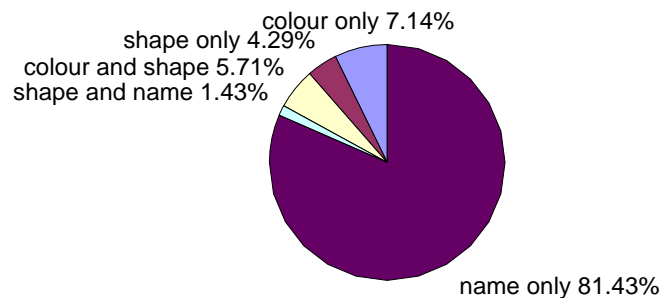


Figure 6. A bar chart showing the relative distribution of hints used in anaphors.

4 Annotating pictures

We propose an annotation scheme here, which we used to annotate the pictures in our corpus. As became clear from our corpus analyses, the least we need to know of the pictures we present as answers is the visual elements they contain, their locations, and some of their basic properties. This corresponds with typical annotations used in image retrieval (i.e. Russell et al., 2005), although it is more extensive. We chose to annotate each visual element with *colour*, *shape*, *function*, *keywords*, and a *noun phrase* uniquely identifying the element. This enables the user to refer to the element using colour, shape, or name, and enables the system to get a noun phrase description of the element for answering visual-element-identity questions. We annotated with a fine granularity, as users sometimes tend to refer to seemingly insignificant elements. We enabled elements to be grouped into aggregates, which can then be labeled as separate elements. The *function* label requires some explanation. It is a type label that describes the purpose of the element in the picture. One would expect that users would usually point to anatomical elements, but we found that they regularly referred to other types of elements. They often encircled label texts, and even asked questions such as 'What does this line mean?' while pointing to a callout line. The least we need to distinguish is between pictures and text labels. In fact, we found that most of the pictures we used are anatomical or histological illustrations, augmented with labels and callouts, and sometimes with flow or transport directions. Other types of

images, such as graphs, charts, and diagrams, were not present in our corpus, as they are less commonly used in the medical texts we used. These would require our current typology to be extended. So, we chose to distinguish the following types: **graphical** (indicating a physical object), **annotation-text** (text labels identifying elements), **callout** (a line connecting a text label with a visual element), and **flow** (typically arrows indicating direction of movement). Some images contain captions embedded in the bitmap. We labeled these as visual elements of the type **caption**, and annotated them with the content of the caption in text. This covered almost all cases, though we occasionally found elements indicating areas or lengths as well. These and everything else was labeled as **other**.

5 Future directions: implementing a dialogue system

We have given an analysis of our corpus of multimodal follow-up questions which should readily help us improve our existing IMIX QA dialogue system. In this section, we shall describe some of the related work in progress. The overall strategy of the dialogue manager is first to detect the utterance type, then react to this by either submitting the question to the QA directly, rewriting it if possible and necessary, or responding with an appropriate prompt to handle special cases. We can add multimodal support in accordance to our findings here by extending the utterance type classifier with the new utterance types proposed here. In particular, the **visual-element-identity** and **visual-property** types proposed in section 2 are completely new types requiring special handling.

We have started work on a classifier that can distinguish between the different types in section 2. As a first experiment, we fed all follow-up questions of the corpus, including the non-multimodal ones, into a machine learning tool, using the number of mouse strokes, and the number of occurrences of specific words and part of speech tags in the sentences as features. Up till now, we have not been able to obtain classification performance above 40%. We did find that particularly important features are the number of gestures in the utterance (with no gestures indicating the question is non-multimodal, and multiple gestures that the question concerns visual discourse), occurrence of the determiner "this" (indicating a regular follow-up question with a deictic anaphor), and occurrence of the word "picture" (indicating visual discourse). To obtain a better performance, we will likely need high-level integrated knowledge, such as dialogue context.

We also tested a gesture recogniser's ability to find the appropriate visual element for encircling gestures. We found that a simple bounding box algorithm, comparing the magnitude of the gesture's bounding box, the visual element's bounding box, and their intersection, could correctly identify 66% of all encircling gestures. A significant part of the failures concerned visual referents which were not annotated and are likely not naturally annotatable (16% of the gestures), and gestures encompassing multiple visual referents in one gesture (6%). Of the remaining referents, our algorithm managed to identify 88% of the referents correctly. We consider this a very good result for such a simple algorithm, showing that resolving gestures' referents is a relatively easy task. We also found that taps, underlines, and arrows need special handling. However, this requires stroke segmentation and gesture type recognition. For this, we plan to adapt from the algorithm in Willems et al. (2005).

A more complex issue our system will need to handle is actual referent detection, and appropriate rewriting or paraphrasing. We first have to detect the appropriate utterance type and anaphor position, and use the labels of the answer's visual elements for further disambiguation. A proper implementation and evaluation of this, and of the resulting dialogue system as a whole, is a matter of future research.

Acknowledgement The IMIX programme is funded by the Netherlands Organisation for Scientific Research (NWO).

References

- Marco De Boni and Suresh Manandhar. 2005. Implementing clarification dialogues in open domain question answering. *Journal of Natural Language Engineering* 11(4):343-361.
- Junichi Fukumoto, Tatsuhiro Niwa, Makoto Itoigawa and Megumi Matsuda. 2004. RitsQA: List answer detection and context task with ellipses handling. Working notes of the Fourth NTCIR Workshop Meeting, Okyo, Japan, pp. 310-314
- Olivier Galibert, Gabriel Illouz and Sophie Rosset. 2005. Ritel: an open-domain, human-computer dialog system. In: *Proceedings of InterSpeech 2005*, Lisbon, Portugal, pp. 909-912.
- Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2005. LabelMe: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025, submitted to Intl. J. Computer Vision.
- Boris van Schooten and Rieks op den Akker. 2006. Follow-up utterances in QA dialogue. Submitted to TAL (Traitement Automatique des Langues/Natural Language Processing), Special Issue on QA.
- Mariet Theune, Boris van Schooten, Rieks op den Akker, Wauter Bosma, Dennis Hof, and Anton Nijholt. 2006. Questions, Pictures, Answers: Introducing Pictures in Question-Answering Systems. In these proceedings.
- Willems, D.J.M., Rossignol, S.Y.P., and Vuurpijl, L.G. 2005. Features for mode detection in natural online pen input. In: *Proceedings of BIGS 2005*, pp. 113-117.