Asking Users About What They Mean: Two Experiments & Results

Hervé Blanchon[a] and Laurel Fais[b]

[a]GETA-CLIPS, BP 53, 38041 Grenoble Cedex, France. herve.blanchon@imag.fr

[b]ATR-ITL, 2-2 Hikaridai, Seika cho, Soraku gun, 619-02 Kyoto, Japan. fais@itl.atr.co.jp

1. INTRODUCTION

1.1 Situation

Natural language (spoken and/or written) is an attractive modality for human-machine interaction. As stated in [8]: speech requires no training; is fast; and requires little attention. Text may be attractive when the utterances are short, when speech is not mandatory, or when speech may be annoying to those surrounding the user. Currently foreseeable applications using a natural language interface include multi-modal drawing tools [3, 7, 11], on-line travel information [5] (and more generally, on-line information retrieval [6]), oral control systems, and finally interpreting communication systems [8, 10].

1.2 Interest

We think that there is a real need to fill the gap between a "toy" and a real-scale application with a component that can overcome difficulties arising while analyzing natural language. Interactive disambiguation of the input is proposed as a reliable solution, which aims to produce more robust, fault-tolerant, and user-friendly software integrated with natural language processing components [1].

The interactive disambiguation process is then a crucial part of the system. How natural and easy it is to use are the preconditions to the success of this idea. There is, thus, a real need to experiment with the design of such a process to be able to propose a wording of the questions used in the disambiguation interaction which will be understandable to users.

1.3 Presentation

What we investigate here is the understandability of the disambiguation dialogues that can be produced by the method described in [2].

In the two experiments for which we give results here, the subjects read a text. They were interrupted to answer questions when an ambiguous sentence was read.

In the next part of this paper we give the results of a pilot experiment and discuss its implications. In the last part, we describe a second experiment and results. In conclusion we draw implications of these results for automatic interactive disambiguation.

2. PILOT EXPERIMENT

2.1 Experimental materials

Two classes of dialogues were used. This required two groups of 12 subjects each. The text and the ambiguities to be solved were the same for each group.

We designed a text that contained a set of 35 ambiguous sentences. The ambiguities in the sentences were selected from naturally occurring ambiguities in a corpus of spontaneous

conversation collected at ATR, Japan [4]. The text itself was made up of two different stories. The 35 ambiguities were distributed evenly over seven categories of ambiguities, i.e., five examples of each category.

Of the five examples in each category, there were two sentences with easy interpretations and three sentences with hard ones. "Easy" interpretations corresponded to the most frequent or most salient interpretation of a given sequence of words and "hard" interpretations corresponded to an unusual, though possible, interpretation of a given sequence of words. The "easy" interpretation tended to be the sense that would "pop up" first in someone's mind for that sequence of words; the hard one was a much less likely interpretation. For example, the easy interpretation of "I want to check in to the hotel" is "I want to register at the hotel;" the hard one is "I want to investigate the hotel."

2.2 Lessons learned

In the course of conducting the pilot experiment and discussing their impressions with subjects afterwards, we learned a number of things that affected the design of the subsequent experiment.

Subjects frequently commented on how unnatural the text seemed to be. There were two reasons why the text sounded unnatural. First, it included actual spoken English examples in written form, surrounded by (made-up) written context. Transcriptions of spoken English often sound unnatural, especially embedded in written text. Second, some of the "hard" interpretations were ones that, in real life, only a computer would have trouble understanding. In trying to motivate these difficult interpretations, unnatural text was produced.

2.3 Recommendations

We made a number of changes to the format for the second experiment based on our experience in the pilot. We realized that using hard interpretations in the pilot experiment was a mistake. It made the text sound unnatural and made the task more difficult for the subjects, clouding the real issue: how well they could respond to the different wordings of the dialogues.

We also changed the arrangement of the screen so that subjects could check back to the text to confirm their understanding of the ambiguity involved. Subjects complained that they could not do this in the pilot; this was also an unnecessary obstacle to the accomplishment of the task.

In the disambiguation interaction, subjects could chose as the correct interpretation, one of two possibilities given in the dialogue box, or they could chose "no answer." While the "no answer" option gave some interesting results, it also made it difficult to see clearly the trends in how subjects answered the questions. For that reason, we designed the next experiment as a forced-choice task.

3. SECOND EXPERIMENT

For this second experiment we chose to ask questions using both a textual mode and spoken one. The spoken mode was added since it seems to be a promising modality [9, 12] in the context of either textual or spoken input (interpreting communication, for example). [Is this OK? I wasn't quite sure I understood this sentence.]

3.1 Setting

For this experiment, the experimenter and the subjects were separated, sitting on either side of a partition. They communicated through head sets (microphone, headphones).

The subject was asked to read aloud, slowly and carefully, a text displayed inside a text window , and pause between each sentence. The scrolling of the text window was controlled by the experimenter (i.e., the text windows of the subject and the experimenter were synchronized).

3.2 Experimental materials

Two classes of dialogues and two modalities were used. This required four groups of subjects; there were fifteen subjects in each group. The text and the ambiguities to be solved were the same for each group.

The text to be read was entirely made up, that is, it didn't contain examples from the "real" corpus as in the pilot experiment. It consisted of three different stories and contained 35 ambiguous sentences.

Two sets of questions were prepared: one human-like, i.e., as if a human were explaining the ambiguities, and one machine-like, i.e., as if the system were generating the explanations, similar to those in the pilot experiment. The contents of the textual and spoken dialogues were the same. The analysis of the results is divided into three parts: statistical analysis, behavioral analysis, and the post-experiment questionnaire analysis.

3.3 Statistical analysis

3.3.1 *Analysis of data actually collected*

If we use the actual answers collected, the basic result is that the difference between the responses to the human-like dialogues and those to the machine-like dialogues is significant ($p<.05$) ; the human-like dialogues were easier to answer. This is affected by two questions in the machine setting which were problematic. We did not present the interpretation choices for these questions consistently with others for the same ambiguities. We will see in another analysis below that the difference is not significant.

It appears that in both machine-like and human-like phrasings, the performance of the subjects tends to be better with text questions, but we can't draw any definitive conclusion since the differences between spoken and textual dialogues are not significant.

3.3.2 *Filtered analysis: problematic questions disregarded*

In this case there is no significant difference between the subjects' performance in the machine-like dialogue settings and in the human-like dialogue settings.

Subjects seem to show better performance for textual dialogues than for spoken dialogues for the human-like phrasings; however, there is no difference at all between text and speech for the machine-like phrasings. Again, the differences are not significant so no definitive conclusion can be drawn.

3.3.3 *Projected analysis: problematic questions corrected*

In this third way of looking at the data, we excluded results for one question from the results for the machine-like dialogues and adjusted the answers to second one according to what we conjecture the answers would have been if the question had been labeled correctly. In this third case, there is again no significant difference between the results in the machine-like and human-like dialogue settings.

In this analysis, the results for spoken and textual dialogues were different for the machine-like and human-like phrasings. The difference is again not significant; thus no definitive conclusion can be drawn.

4. CONCLUSION AND PERSPECTIVES

If we allow for the problematic questions we see that there were no significant differences according to the style (machine, or human) of the presentation of the disambiguation dialogues, and no significant differences according to the modality (spoken or textual). The former result is essential to the success of an automatic interactive disambiguation program. We have seen that subjects are able to interpret the dialogues when presented in human-like, i.e., natural, phrasing, but it is not likely that automatically generated dialogues can be so natural. Therefore, it is critical that users be able to interpret the type of dialogues that machines are likely to be able to generate. The results reported here show that this is indeed the case.

We also investigated whether spoken or textual dialogues would be easier to understand. This is a design question; it affects how an automatic system will be designed, but is not crucial to the system. The results found here, as well as comments made by some of the participants about wanting to have text instead of speech, suggest that one design feature for an interactive disambiguation system should be the option for users to choose in which modality they would

like to have the dialogues presented. According to our results, both modalities are understandable.

Although the “repeat” option was not extensively used in the spoken setting, it is still necessary to include it for cases where users cannot understand the dialogue after the first hearing. Other suggestions made by the subjects can be easily implemented. For example, more of the context of the ambiguity can be included in the dialogue; this would also support the most frequent strategy used by the subjects in determining their responses, i.e., the use of context. In addition, spoken utterances can be made shorter and faster. How best to use intonation in the spoken presentation of disambiguation dialogues is an open and interesting question.

It will be also necessary to run an experiment using as the text, one provided by the subjects themselves. This may be the only way to have a better analysis of the interactive disambiguation methodology we have proposed.

REFERENCES

[1] Blanchon, H. Clarification: Towards more User-Friendly Natural Language Human-Computer Interaction. in Proceedings of Poster session of HCI'95, 6th International Conference on Human-Computer Interaction. (Yokohama, Japan, July 9-14, 1995), vol. 1/1 : 42-42.

[2] Blanchon, H. Interactive Disambiguation of Natural Language Input: a Methodology and two Implementations for French & English. in Proceedings of IJCAI-97. 23-29 August, 1997) : to be published.

[3] Caelen, J. Multimodal Human-Computer Interaction. in Fundamentals of Speech Synthesis and Speech Recognition. John Wiley & Sons. New York. 1994.

[4] Fais, L. and Blanchon, H. Ambiguities in Task-oriented Dialogues. in Proceedings of MIDDIM'96. (Le col de porte, Isère, France, 12-14 Août 1996), vol. 1/1 : 263-275.

[5] Goddeau, D., Brill, E., Glass, J., Pao, C., Philips, M., Polifroni, J., Seneff, S. and Zue, V. GALAXY: a Human-Language Interface to On-Line Travel Information. in Proceedings of ICSLP 94. (Yokohama, Japan, September 18-22, 1994), vol. 2/4 : 707-710.

[6] Haddock, N. J. Multimodal Database Query. in Proceedings of Coling-92. (Nantes, France, 23-28 juillet 1992), vol. 4/4 : 1274-1278.

[7] Hiyoshi, M. and Shimazu, H. Drawing Pictures with Natural Language and Direct Manipulation. in Proceedings of Coling-94. (Kyoto, Japan, August 5-9, 1994), vol. 2/2 : 722-726.

[8] Kay, M., Gawron, J. M. and Norvig, P. Verbmobil: A Translation System for Face-to-Face Dialog. CSLI lecture note no 33. Center for the Study of Language and Information, Stanford, CA. 1994.

[9] Lehiste, I., Olive, J. P. and Streeter, L. A. Role of duration in disambiguating syntactically ambiguous sentences. Journal of the Acoustical Society of America. 60, 5, 1199-1202.

[10] Morimoto, T., Suzuki, M., Takezawa, T., Kikui, G. i., Nagata, M. and Tomokiyo, M. A Spoken Language Translation System: SL-TRANS2. in Proceedings of Coling-92. (Nantes, France, 23-28 juillet 1992), vol. 3/4 : 1048-1052.

[11] Nishimoto, T., Shida, N., Kobayashi, T. and Shirai, K. Multimodal Drawing Tool Using Speech, Mouse and Key-Board. in Proceedings of ICSLP 94. (Yokohama, Japan, September 18-22, 1994), vol. 3/4 : 1287-1290.

[12] O'Shaughnessy, D. Specifying accent marks in French text for teletext and speech synthesis. International Journal of Man-Machine Studies. 31, 4, 405-414.